PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-231238

(43)Date of publication of application : 05.09.1997

(51)Int.Cl. G06F 17/30

(21)Application number : 08-058391

(22)Date of filing : 20.02.1996

(71)Applicant : OMRON CORP

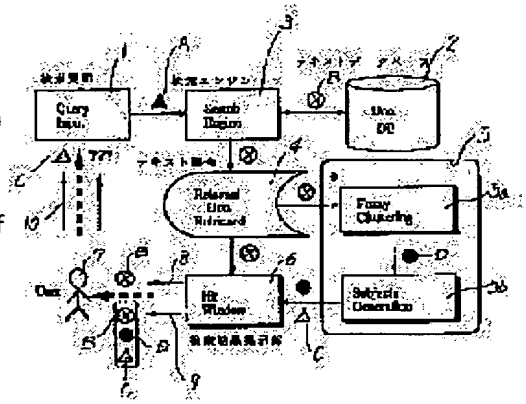
(72)Inventor : GO ATOU
SOGO TAIJI
SAWADA AKIRA

(54) DISPLAY METHOD FOR TEXT RETRIEVAL RESULT AND DEVICE THEREFOR

(57)Abstract:

PROBLEM TO BE SOLVED: To improve both retrieval efficiency and accuracy by dividing a text set into plural groups based on the theme analysis result of every text, generates the theme sort information showing the attribute of every group, and displays these information in every group.

SOLUTION: A retrieval engine 3 expands a retrieval expression based on a prescribed algorithm and extracts a relative text set 4 out of a document data base 2. A fuzzy gathering part 5a of a processing part 5 divides the set 4 into plural groups based on the theme analysis result of every text, and a theme sort information generation part 5b generates the theme sort information showing the attribute of every group. A retrieval result display part 6 processes the acquired information (text set B, centroid D and theme sort information C) in a prescribed display mode and shows them to a user 7. As a result, the document retrieval result can be easily confirmed and the retrieval efficiency and accuracy can be improved owing to prevention of the retrieval omission.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

CLAIMS

[Claim(s)]
[Claim 1] The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, The generation step which generates the theme classification information expressing the attribute of groups involved about each of each group which was obtained by said division step, The text browsing result method of presentation characterized by what the display step which classifies and displays the theme classification information of each group who asked at said generation step according to a group is provided for.
[Claim 2] The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, The generation step which generates the theme classification information expressing the attribute of groups involved about each of each group which was obtained by said division step, The group goodness of fit calculation step which asks for the goodness of fit between the group and said retrieval conditions about each of each of said group, The text browsing result method of presentation characterized by what the display step which classifies the theme analysis information of each group who asked at said generation step according to a group, and displays it on descending of the goodness of fit for which it asked by said goodness of fit calculation step is provided for.
[Claim 3] The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, Based on the analysis result of the contents of each text in said group, whenever [affiliation / which computes whenever / affiliation / to the groups involved of each text] a calculation step and in said two or more groups The text browsing result method of presentation characterized by what the selection step for choosing the group who becomes a text display object, and the display step which indicates the text in the group chosen at said selection step by contents at the order of whenever [said affiliation / which was computed] are provided for.
[Claim 4] The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, Based on the analysis result of the contents of each text in said group, the goodness of fit calculation step which computes the goodness of fit to said retrieval conditions of each text, and in said two or more groups The text browsing result method of presentation characterized by what the selection step for choosing the group who becomes a text display object, and the display step which indicates the text in the group chosen at said selection step by contents at the order of said computed goodness of fit are provided for.
[Claim 5] The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, Based on the analysis result of the contents of each text in said group, whenever [affiliation / which computes whenever / affiliation / to the groups involved of each text] A calculation step, Based on the analysis result of the contents of each text in said group, the goodness of fit calculation step which computes the goodness of fit to said retrieval conditions of each text, and in said two or more

groups The group selection step for a display for choosing the group who becomes a text display object, The display-order criteria selection means for choosing whether the text in said each group is displayed in order of the goodness of fit to retrieval conditions, or it displays in order of whenever [to groups involved / affiliation], The text browsing result method of presentation characterized by what the display step which indicates the text in the group chosen at said group selection step for a display by contents at the order of the display-order criteria chosen with said display-order criteria selection means is provided for.

[Claim 6] Said division step is the text browsing result method of presentation according to claim 1 to 5 characterized by what the text set obtained by searching a database based on the given retrieval conditions is divided into two or more groups for using the fuzzy clustering method.

[Claim 7] The theme classification information expressing the attribute of the groups involved generated at said generation step is the text browsing result method of presentation given in either claim 1 characterized by what is been what expresses the attribute of groups involved by the group of some keywords, or claim 2.

[Claim 8] The theme classification information expressing the attribute of the groups involved generated at said generation step is the text browsing result method of presentation given in either claim 1 characterized by what is been what expresses the attribute of the section loop formation concerned by the short text, or claim 2.

[Claim 9] Whenever [affiliation / which performs fuzzy clustering to the description matrix of the text set obtained by searching a database based on the given retrieval conditions, and generates whenever / to each classification category / affiliation / for every document] A generation step, Using whenever [said affiliation / which was generated] the document allotment step which assigns each document to 1 or two or more classification categories, and in said two or more classification categories The classification category selection step for choosing the classification category used as a text display object, The text browsing result method of presentation characterized by what the display step which indicates the text in the classification category chosen at said classification category selection step by contents at the order of a goodness of fit to the group is provided for.

[Claim 10] Said document allotment step is the text browsing result method of presentation according to claim 9 characterized by what each document is assigned for to the classification category of k high orders of whenever [affiliation].

[Claim 11] Said document allotment step is the text browsing result method of presentation according to claim 9 characterized by what each document is assigned for to the classification category which has a value whenever [beyond a certain threshold alpha / affiliation].

[Claim 12] Said document allotment step is the text browsing result method of presentation according to claim 9 characterized by what each document is assigned for to a classification category in consideration of the probability distribution of a category.

[Claim 13] A division means to divide automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, A generation means to generate the theme classification information expressing the attribute of groups involved about each of each group which was obtained by said division means, The text browsing result display characterized by what a display means to classify and display the theme classification information of each group who asked with said generation means according to a group is provided for.

[Claim 14] A division means to divide automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, A generation means to generate the theme classification information expressing the attribute of groups involved about each of each group which was obtained by said division means, A group goodness of fit calculation means to ask for the goodness of fit between the group and said retrieval conditions about each of each of said group, The text browsing result display characterized by what a display means to classify the theme analysis information of each group who asked with said generation means according

to a group, and to display it on descending of the goodness of fit for which it asked with said goodness of fit calculation means is provided for.

[Claim 15] A division means to divide automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, Based on the analysis result of the contents of each text in said group, whenever [affiliation / which computes whenever / affiliation / to the groups involved of each text] a calculation means and in said two or more groups The text browsing result display characterized by what the selection means for choosing the group who becomes a text display object, and the display means which indicates the text in the group chosen with said selection means by contents at the order of whenever [said affiliation / which was computed] are provided for.

[Claim 16] A division means to divide automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, Based on the analysis result of the contents of each text in said group, a goodness of fit calculation means to compute the goodness of fit to said retrieval conditions of each text, and in said two or more groups The text browsing result display characterized by what the selection means for choosing the group who becomes a text display object, and the display means which indicates the text in the group chosen with said selection means by contents at the order of said computed goodness of fit are provided for.

[Claim 17] A division means to divide automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text, Based on the analysis result of the contents of each text in said group, whenever [affiliation / which computes whenever / affiliation / to the groups involved of each text] A calculation means, Based on the analysis result of the contents of each text in said group, a goodness of fit calculation means to compute the goodness of fit to said retrieval conditions of each text, and in said two or more groups The group selection means for a display for choosing the group who becomes a text display object, The display-order criteria selection means for choosing whether the text in said each group is displayed in order of the goodness of fit to retrieval conditions, or it displays in order of whenever [to groups involved / affiliation], The text browsing result display characterized by what the display means which indicates the text in the group chosen with said group selection means for a display by contents at the order of the display-order criteria chosen with said display-order criteria selection means is provided for.

[Claim 18] Said division means is a text browsing result display according to claim 13 to 17 characterized by what the text set obtained by searching a database based on the given retrieval conditions is divided into two or more groups for using the fuzzy clustering method.

[Claim 19] The theme classification information expressing the attribute of the groups involved generated with said generation means is a text browsing result display given in either claim 13 characterized by what is been what expresses the attribute of groups involved by the group of some keywords, or claim 14.

[Claim 20] The theme classification information expressing the attribute of the groups involved generated with said generation means is a text browsing result display given in either claim 13 characterized by what is been what expresses the attribute of groups involved by the short text, or claim 14.

[Claim 21] Whenever [affiliation / which performs fuzzy clustering to the description matrix of the text set obtained by searching a database based on the given retrieval conditions and generates whenever / to each classification category / affiliation / for every document] A generation means, Using whenever [said affiliation / which was generated] a document allotment means to assign each document to 1 or two or more classification categories, and in said two or more classification categories The classification category selection means for choosing the classification category used as a text display object, The text browsing result display characterized by what the display means which indicates the text in the classification category chosen with said classification category selection means by contents at the

order of a goodness of fit to the group is provided for.
[Claim 22] Said document allotment means is a text browsing result display according to claim 21 characterized by what each document is assigned for to the classification category of k high orders of whenever [affiliation].
[Claim 23] Said document allotment means is a text browsing result display according to claim 21 characterized by what each document is assigned for to the classification category which has a value whenever [beyond a certain threshold alpha / affiliation].
[Claim 24] Said document allotment means is a text browsing result display according to claim 21 characterized by what each document is assigned for to a classification category in consideration of the probability distribution of a category.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention] This invention relates to the suitable text-browsing result method of presentation and the equipment for retrieval of a document database, divides automatically into two or more groups the text set obtained by searching a database based on the retrieval conditions given especially using the theme analysis result of each text, and relates to the text-browsing result method of presentation and the equipment which classify the theme classification information expressing each group's attribute acquired by this division according to a group, and displayed it.
[0002]
[Description of the Prior Art] As a conventional text browsing result display, what was indicated by JP,6-76004,A is known, for example.
[0003] The database retrieval solution storing section in which this equipment stores a database retrieval result, A distance calculation means between retrieval solutions to compute the distance between each retrieval solution by seasoning with a user's control input two or more attribute value which said database retrieval solution has, A retrieval solution group division means to divide into the group of the number which specified the retrieval solution as the user using the distance between retrieval solutions, or the number defined beforehand, A group representation retrieval solution calculation means to compute the retrieval solution located near an affiliation group's center of gravity, It consists of a representation retrieval solution selection means as which a user is made to choose a specific retrieval solution out of each group's representation retrieval solution, and a retrieval solution display means in a group to display all the retrieval solutions in the group to whom the representation retrieval solution belongs.
[0004] Namely, if it is in equipment conventionally [this], it is what classifies the database (numeric value) retrieval solution structured by the non-overlapping technique into the number of classifications

which the user specified. It displays without carrying out ranking of all the retrieval solutions in the group who was made to choose the group who displays one retrieval solution nearest to a group's classified center of gravity at a time as pilot data, and expects it of a user (with no ranking), and was chosen.

[0005]

[Problem(s) to be Solved by the Invention] However, if shown in such a conventional text browsing result display (retrieval solution display), there was a trouble said that application in a destructuring database like the full text is difficult for the following reason.

[0006] That is, when a representation retrieval solution is a representation document in a group since the representation retrieval solution of the center-of-gravity location in a group is displayed if it is in equipment such conventionally, since what expresses the contents of the representation document directly is not displayed but the whole document is displayed, it is hard to grasp the group's contents. That is, in order to show each group's classified theme semantics, it is desirable for there to be a specific past ** case in contents, to extract the attribute item group [-like in common] in a group rather, and to show a user only by display one retrieval solution nearest to a mere group's center of gravity at a time as pilot data. In addition, if it is in the case of a full-text search system, it is meaningless to show all attribute data as they are as pilot data, and a new definition of pilot data which can understand the contents of a document easily is desired.

[0007] Moreover, if it is in equipment conventionally, since a group is not arranged in order of the goodness of fit to retrieval conditions, it is hard to choose the group who agreed for the purpose of retrieval. In addition, only by referring to the representation solution to, if it is in equipment conventionally, since the solution in a group is not located in a line in order of whenever [to a group / affiliation], even when it is hard to grasp a group's image, it is difficult [it] to grasp an image with reference to other solutions. That is, by the method displayed without carrying out ranking of all the retrieval solutions in the selected group, if the classification number of cases increases, a user's burden will benefit specification large to a retrieval result. In order to mitigate such a burden and to raise retrieval effectiveness, a ranking function which can promote specification in a retrieval result is desired.

[0008] Furthermore, since usually has two or more themes, a document has a possibility of producing leakage in a retrieval result on the display of a document classification result by the conventional technique of classifying one document only into one cluster. Therefore, in case a theme classification is performed to a document-retrieval result set, the overlapping technique which allows belonging to the cluster from which plurality differs (the theme is expressed) is desired.

[0009] The place which this invention is made in view of an above-mentioned trouble, and is made into that purpose The check over a document-retrieval result is made easy. In improvement in retrieval effectiveness, and a list Can aim at improvement in the retrieval precision by prevention of the leakage in retrieval, and it also becomes the guide of how moreover, the shown theme information extracts data efficiently and puts them. It is in offering the retrieval result method of presentation and equipment which enabled it to perform advanced adaptation retrieval (Relevance Feedback) using this added response indication.

[0010]

[Means for Solving the Problem] Invention of a publication to claim 1 (or claim 13) of this application The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text (or means), The generation step which generates the theme classification information expressing the attribute of groups involved about each of each group which was obtained by said division step (or means) (or means), It is shown in the text browsing result method of presentation (or equipment) characterized by what the display step (or means) which classifies and displays the theme classification information of each group who asked at said generation step (or means) according to a group is provided for.

[0011] Here, the text set of the homepage which exists on the text set remembered to be a "database" by mass storage media, such as a hard disk and an optical disk, or the Internet is equivalent to this.
[0012] Moreover, "theme analysis" means generating the information which shows the contents of the text directly, and the set of the keyword on the title in a document may be generated. In the gestalt of operation, the vector (Fi) which is expressing the document by the feature vector in document space is equivalent to this.
[0013] Moreover, "theme classification information" means the information which shows the group's contents directly about the group of a text. Two methods of a keyword method and a text method are shown by the gestalt of operation.
[0014] And since according to invention of this claim 1 (or claim 13) the information which expresses a group directly is added and it indicates by the partition according to a group, it becomes easy to grasp the overview of the group who constitutes a retrieval result, and the group selection for the next processing becomes very easy.
[0015] Invention of claim 2 (or claim 14) of this application The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text (or means), The generation step which generates the theme classification information expressing the attribute of groups involved about each of each group which was obtained by said division step (or means) (or means), The group goodness of fit calculation step which asks for the goodness of fit between the group and said retrieval conditions about each of each of said group (or means), The theme analysis information of each group who asked at said generation step (or means) It is shown in the text browsing result method of presentation (or equipment) characterized by what the display step (or means) which classifies according to a group and is displayed on descending of the goodness of fit for which it asked by said goodness of fit calculation step is provided for.
[0016] And according to invention of this claim 2 (or claim 14), since it displays on said claim 1 (or claim 13) in order of the goodness of fit to retrieval conditions in addition to the effect of the invention of a publication, the group who agreed for the purpose of retrieval can be chosen, checking the group's contents.
[0017] Invention of claim 3 (or claim 15) of this application The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text (or means), Based on the analysis result of the contents of each text in said group, whenever [affiliation / which computes whenever / affiliation / to the groups involved of each text] a calculation step (or means) and in said two or more groups The selection step for choosing the group who becomes a text display object (or means), It is shown in the text browsing result method of presentation (or equipment) characterized by what the display step (or means) which indicates the text in the group chosen at said selection step (or means) by contents at the order of whenever [said affiliation / which was computed] is provided for.
[0018] And since the text in the selected group is displayed in order of whenever [to a group / affiliation] according to invention of this claim 3 (or claim 15), it becomes easy to grasp a definition of a group.
[0019] Invention of claim 4 (or claim 16) of this application The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text (or means), Based on the analysis result of the contents of each text in said group, the goodness of fit calculation step (or means) which computes the goodness of fit to said retrieval conditions of each text, and in said two or more groups The selection step for choosing the group who becomes a text display object (or means), It is shown in the text browsing result method of presentation (or equipment) characterized by what the display step (or means) which indicates the text in the group chosen at said selection step (or means) by contents at the order of said computed goodness of fit is provided for.

[0020] And since according to invention of this claim 4 (or claim 16) the group suitable for retrieval conditions is chosen and the text in it is further displayed in order of retrieval conditions, a suitable text is early displayed certainly to retrieval conditions rather than the case where a text is displayed in order of a goodness of fit without carrying out the group division of the retrieval result.
[0021] Invention of claim 5 (or claim 17) of this application The division step which divides automatically the text set obtained by searching a database based on the given retrieval conditions into two or more groups using the theme analysis result of each text (or means), Based on the analysis result of the contents of each text in said group, whenever [affiliation / which computes whenever / affiliation / to the groups involved of each text] A calculation step (or means), Based on the analysis result of the contents of each text in said group, the goodness of fit calculation step (or means) which computes the goodness of fit to said retrieval conditions of each text, and in said two or more groups The group selection step for a display for choosing the group who becomes a text display object (or means), The display-order criteria selection step for choosing whether the text in said each group is displayed in order of the goodness of fit to retrieval conditions, or it displays in order of whenever [to groups involved / affiliation] (or means), It is shown in the text browsing result method of presentation (or equipment) characterized by what the display step (or means) which indicates the text in the group chosen at said group selection step for a display by contents at the order of the display-order criteria chosen with said display-order criteria selection means is provided for.
[0022] And according to invention of this claim 5 (or claim 17), according to a user's purpose, the display order of a text is changeable.
[0023] Invention of a publication to claim 6 (or claim 18) of this application In the text browsing result method of presentation (or equipment) given in either claim 1 (or claim 13) thru/or claim 5 (or claim 17) said said division step (or means) It is characterized by what the text set obtained by searching a database based on the given retrieval conditions is divided into two or more groups for using the fuzzy clustering method.
[0024] And since the fuzzy classification (theme classification) by the contents of the theme is automatically performed by the orverlapping method to the document set discovered by a certain retrieval type according to invention given in this claim 6 (or claim 18), improvement in the retrieval precision by prevention of the leakage in retrieval is expectable.
[0025] Theme classification information expressing the attribute of the groups involved from which invention of a publication is generated by claim 1 (or claim 13) or claim 2 (or claim 14) at said generation step (or means) in the text browsing result method of presentation (or equipment) of a publication at claim 7 (or claim 19) of this application is characterized by what is been what expresses the attribute of groups involved by the group of some keywords.
[0026] And according to invention given in this claim 7 (or claim 19), the attribute of groups involved can be intuitively grasped through the group of some keywords.
[0027] Theme classification information expressing the attribute of the groups involved from which invention of a publication is generated by claim 1 (or claim 13) or claim 2 (or claim 14) at said generation step (or means) in the text browsing result method of presentation (or equipment) of a publication at claim 8 (or claim 20) of this application is characterized by being what expresses the attribute of groups involved by the short text.
[0028] And according to invention given in this claim 8 (or claim 20), the attribute of groups involved can be made to understand intelligibly for anyone through a short text.
[0029] Invention of a publication to claim 9 (or claim 21) of this application Fuzzy clustering is performed to the description matrix of the text set obtained by searching a database based on the given retrieval conditions. Whenever [affiliation / which generates whenever / to each classification category / affiliation / for every document] A generation step (or means), Using whenever [said affiliation / which was generated] the document allotment step (or means) which assigns each document to 1 or two or more classification categories, and in said two or more classification categories The classification

category selection step for choosing the classification category used as a text display object (or means), It is shown in the text browsing result method of presentation (or equipment) characterized by what the display step (or means) which indicates the text in the classification category chosen at said classification category selection step (or means) by contents at the order of a goodness of fit to the group is provided for.

[0030] And according to invention given in this claim 9 (or claim 21), each document is assigned to 1 or two or more classification categories using the overlapping technique, and since the text in the classification category chosen in that condition is indicated by contents at the order of a goodness of fit to that group, improvement in the retrieval precision by prevention of the leakage in retrieval can be aimed at in improvement in retrieval effectiveness, and a list.

[0031] Invention given in claim 10 (or claim 22) of this application is characterized by what said document allotment step (or means) assigns each document for in the classification category of k high orders of whenever [that affiliation] in invention given in said claim 9 (or claim 21).

[0032] And according to invention given in this claim 10 (or claim 22), the high order of whenever [affiliation] can be made to always display the document of the fixed number on claim 9 (or claim 21) about each classification category in addition to the effect of the invention of a publication.

[0033] Invention given in claim 11 (or claim 23) of this application is characterized by what said document allotment step assigns each document for to the classification category which has a value whenever [beyond a certain threshold alpha / affiliation] in invention given in said claim 9 (or claim 21).

[0034] And according to invention given in this claim 11 (or claim 23), the document which has a value whenever [beyond the threshold alpha which is always in claim 9 (or claim 21) about each classification category in addition to the effect of the invention of a publication / affiliation] can be displayed.

[0035] Invention given in claim 12 (or claim 24) of this application is characterized by what said document allotment step assigns each document for in a classification category in consideration of the probability distribution of a category in invention given in said claim 9 (or claim 21).

[0036] And according to invention given in this claim 12 (or claim 24), the document which relates to claim 9 (or claim 21) in consideration of the probability distribution of a category about each classification category in addition to the effect of the invention of a publication can be displayed.

[0037]

[Embodiment of the Invention] Below, the gestalt of suitable operation of this invention approach and equipment is explained at a detail, referring to an accompanying drawing.

[0038] First, the functional block diagram of drawing 1 shows notionally the configuration of the text browsing equipment with which this invention approach and equipment were applied. In this drawing, 1 is the retrieval question input section (it is described as Query Input) for inputting the feedback retrieval question (FeedBack Query) which should be inputted at the time of the original retrieval question (Original Query) and retrieval narrowing down which should be inputted at the time of retrieval initiation, and, specifically, is constituted as everyone knows by the software for an input processed in control units and those signals, such as a mouse and a keyboard.

[0039] 2 is a text (document) database (it is described as Doc.DB) equivalent to the text set used as the candidate for retrieval, and the text set of the homepage which specifically exists on the text set memorized by mass storage media, such as a hard disk and an optical disk, or the Internet is equivalent to this.

[0040] 3 is a search engine (it is describe as Search Engine) locate in the center of a text retrieval system , according to a predetermined algorithm , a retrieval type is develop concrete as everyone knows by making into retrieval conditions the original retrieval question (Original Query) and the feedback retrieval question (FeedBack Query) which are input from the above-mentioned retrieval question input section 1 , and the software to extract corresponds to this the text set relate from the above-mentioned document database 2 .

[0041] 4 is the related text set (it is described as Relevant Doc.Retrieved) which did in this way and was

extracted by the search engine (Search Engine) 3, and this text set 4 is set as the object of the processing processing in this invention so that it may mention later.

[0042] Fuzzy grouping section (it is described as Fuzzy Clustering) 5a equivalent to a division means by which 5 is the processing processing section equivalent to the important section of this invention, and this processing processing section 5 divides the text set 4 into two or more groups automatically using the theme analysis result of each text, In this way, theme classification information generation section (it is described as Subject Generation) 5b which generates the theme classification information expressing the attribute of groups involved about each of each group which was obtained is constituted as a core.

[0043] An operation of fuzzy grouping section (Fuzzy Clustering) 5a and theme classification information generation section (Subject Generation) 5b is notionally shown in drawing 2 . In this drawing, the field surrounded as the continuous line shown with a sign 4 expresses the whole text set (Relevant Doc.Retrieved) extracted by the search engine (Search Engine) 3.

[0044] Three fields which similarly were surrounded with the broken line shown with Signs 4a, 4b, and 4c express each of three groups which was divided in the fuzzy grouping section (Fuzzy Clustering) 5.

[0045] The black painting trigonum mark shown with Sign A expresses the original retrieval question (Original Query) inputted at the time of retrieval initiation. The round mark containing x shown with Sign B expresses each of each configuration text of the text set 4 in which the retrieval extract was carried out by the input of the original retrieval question (Original Query) A.

[0046] Three void trigonum marks shown with Signs calcium, Cb, and Cc express the theme classification information (Group Subject) expressing Groups' 4a, 4b, and 4c attribute. In addition, such theme classification information calcium, Cb, and Cc is suitable also as a feedback retrieval question (FeedBack Query), if it uses for retrieval narrowing-down ********.

[0047] Three black painting round marks shown with Signs Da, Db, and Dc express Groups' 4a, 4b, and 4c center of gravity. Similarly, the black painting square mark shown with Sign D expresses the center of gravity of the text set 4.

[0048] Fuzzy grouping section (Fuzzy Clustering) 5a divides the text set 4 into the groups [two or more (this example three pieces)] 4a, 4b, and 4c by performing well-known fuzzy clustering processing to the text set 4 obtained as a result of retrieval so that clearly from drawing 2 . On the other hand, theme classification information generation section (Subject Generation) 5b generates the theme classification information calcium, Cb, and Cc expressing the attribute of groups involved about each of each groups 4a, 4b, and 4c which was obtained in this way. The theme classification information calcium, Cb, and Cc expressing the attribute of the groups involved acquired by doing in this way differs in each groups' 4a, 4b, and 4c centers of gravity Da, Db, and Dc, and becomes what expressed each group's attribute directly surely so that clearly from drawing. In addition, the contents of processing of such fuzzy grouping section (Fuzzy Clustering) 5a and theme classification information generation section (Subject Generation) 5b are explained in more detail later.

[0049] the information (the text set B, a center of gravity D, theme classification information C) which it returns to drawing 1 , and 6 is the retrieval result presentation section (it is described as Hit Window) which is equivalent to the important section of this invention similarly, and was acquired by the above-mentioned progress in this retrieval result presentation section (Hit Window) 6 -- predetermined display voice -- after processing it like, it shows to a user (it is described as User) 7. Those display modes are also later explained to a detail.

[0050] In addition, in drawing 1 , the information flow by equipment is shown in coincidence conventionally which was expressed by the information flow and broken line by this invention expressed by the continuous line. Namely, in the retrieval result presentation section (Hit Window) 6, if it is in equipment conventionally, as shown in the broken-line arrow head 8, when the number of texts which it is only showing a user 7 as it is the text set B obtained as a result of retrieval, and is contained in the text set B in this case is abundant, as for a user, inconvenience is caused to discovering the target text.

If it is in this invention, on the other hand, in the retrieval result presentation section (Hit Window) 6 As shown in the continuous-line arrow head 9, only not only in the text set B obtained as a result of retrieval Since a user 7 will be shown even the theme classification information (Group Subject) C at the center-of-gravity (Clustercentroids) D list of each classification, Especially, it becomes possible by making this theme classification information (Group Subject) C into a key to discover the target text easily. Namely, as shown in the continuous-line arrow head 10 The acquired theme classification information C (it is equivalent to C1, C2, and C3 of drawing 2) thus, as it is If it gives the retrieval question input section (Query Input) 1 as a feedback retrieval question (FeedBack Query) C (signs that a retrieval question branches by the continuous-line arrow head 11 of drawing 2 R> 2 "Query Splitting" are shown), the text set 4 will be narrowed down exactly. The target text can be discovered easily, namely, advanced adaptation retrieval (relevance feedback) can be made to perform.

[0051] Next, the text browsing equipment explained notionally above is explained to a detail with reference to the drawing below drawing 3 focusing on data processing for realizing the screen-display mode and it further.

[0052] The whole data processing in the text browsing equipment concerning this invention is shown in the General flow chart of drawing 3 . In addition, the processing shown in this General flow chart is started by choosing one of the menu item of that in a predetermined system menu.

[0053] If processing is started in this drawing, a retrieval screen will be displayed by the predetermined display mode on the screen of the image display machine which constitutes retrieval equipment (step 301). Thus, an example of the retrieval screen displayed is shown in drawing 4 . As shown in this drawing, the display screen is constituted by the longwise rectangle-like window W1, and the part of the up abbreviation 1/3 is made into the retrieval question input area A1, and let the part of the lower abbreviation 2/3 be the retrieval result output area A2.

[0054] Into the retrieval question input area A1, the window W2 for a retrieval question input prepares, and is carried out. To this window W2 up side the input guide sentence (Enter Query in plain English) 12 -- moreover, in the right-hand side The start button 13 for giving the starting command to the search engine (Search Engine) 3 mentioned above (it is described as the inside O.K. of drawing), The cancellation carbon button (it is described as the inside CANCEL of drawing) 14 for canceling a retrieval question (Query) and the help button (it is described as the inside HELP of drawing) 15 for asking for actuation exchange etc. from a system are formed.

[0055] In the retrieval result output area A2, window W3 for a retrieval result output is prepared, and the scroll bar 16 is formed in the right-hand side of this window W3. furthermore, to this retrieval result output area A2 down side The whole sentence demand carbon button 17 for requiring a text whole sentence output as a retrieval result (it is described as Full Text in drawing), The QBE carbon button 18 and the classification-ized demand carbon button 19 for requiring classification-ization of a retrieval result (it is described as the inside Grouping of drawing), The abstract demand carbon button (it is described as the inside Summarize of drawing) 20 for requiring a text abstract output as a retrieval result and the reset button (it is described as the inside Back of drawing) 21 for returning a screen to the initial output state of a retrieval result are formed.

[0056] In addition, it cannot be overemphasized that it is performed by click actuation of a mouse etc. after actuation of various kinds of above carbon buttons 13, 14, 15, 16, 17, 18, 19, 20, and 21 moves cursor to the carbon button of hope.

[0057] And when a retrieval question is inputted like "I wantto know Clinton's political condition." with a natural language (especially this example English) from a keyboard according to the input guide sentence (Enter Query in plain English) 12, this inputted retrieval question 22 will be displayed in a window W2.

[0058] In this condition, if a start button (it is described as the inside O.K. of drawing) 13 is operated Return to drawing 3 , retrieval/display process is performed, and a search engine (Search Engine) 3 is started. The text set 4 relevant to a retrieval question is extracted from the document database 2, each configuration text of the extracted text set is sorted by this order with a high goodness of fit with the

retrieval question 22, and only that title 23 is displayed in window W3 (step 302). Moreover, the goodness of fit marks 24a, 24b, and 24c which classify the goodness of fit to the retrieval question of the text concerned into three steps ("quantity", "inside", low ["low"]), and express it are displayed on the head part of the title 23 of each text. Here, goodness of fit mark 24c goodness of fit mark 24b goodness of fit mark 24a shown by the round mark of black painting out is indicated to be to a goodness of fit "quantity" by the round mark of gray painting out is indicated to be to a goodness of fit "inside" by the round mark of void supports the goodness of fit "low", respectively.

[0059] Henceforth, it returns to drawing 3 and will be in the condition of standing by selection of a text processing facility to a system side (step 303). In this condition, actuation of the classification-ized demand carbon button (Grouping) 19 shown in the screen of drawing 4 performs classification-ized processing which is the important section of this invention (step 306).

[0060] The detail of classification-ized processing is shown in drawing 5 . If processing is started in this drawing, it will be in the condition of standing by assignment of classification group number g, by showing a predetermined initial screen format (step 501). In this condition, completion of assignment (this example "5") of classification group number g performs in order extract processing (step 502) of the document characteristic quantity which is the description part of this invention, fuzzy clustering processing (it is described as Fuzzy Clustering) (step 503), and generation processing (step 504) of theme classification information.

[0061] In extract processing (step 502) of document characteristic quantity, generation of document abstraction and a document feature vector is performed as follows. A document is expressed by the set (vector which uses a word as a component) of eclipse **** with weight, and the set of a document is expressed as a matrix which uses a word as a component. Therefore, a document is abstracted like several 1 by starting automatically the word (important word) used as the description of each document, making the class of word into Dimension m, and using the vectorial representation Fi to which each element is proportional to the frequency of occurrence of the word of a document unit.

[0062]

[Equation 1]

$$Fi=(fi1, fi2, ..., fim)$$

Ｆｉ：文書ｉの特徴ベクトル
ｆｉｊ：単語ｊの文書ｉに対する重み
（頻度、或は他の評価値）

数１

The example of a document vector set is shown in Table 1. The weight (for example, frequency) of the important word (Clinton, Singapore, China--) included in each of the configuration document (F1, F2, F3 --) of a document set is shown by this example.

[0063]

[Table 1]

	Clinton	Singapore	China ...
F1	0.8	0.4	0.0
F2	0.6	0.7	0.0
F3	0.5	0.0	0.7
...			

文書ベクトル集合の例

表１

The example which developed the document vector set shown in Table 1 to document space is shown in drawing 6 . In this example, each configuration document (F1, F2, F3 --) of a document set is developed by the document space which sets an axis of coordinates as the above-mentioned important word (Clinton, Singapore, China--).

[0064] the description matrix of the document set as a retrieval result in the continuing fuzzy clustering processing (step 503) -- receiving -- well-known FCM -- two kinds of classification information (Vc, Ui) as follows is generated by performing fuzzy clustering using law.

[0065] 1) the representation document feature vector Vc of each classification -- [Equation 2]

$$Vc=(vc1, vc2, ..., vcm)$$

Ｖｇ：分類グループｇの代表文書ベクトル
ｖｃｊ：単語ｊの分類グループｃの代表文書に対する重み

数 2

2) whenever [to each classification category of each document / affiliation] -- Ui -- [Equation 3]

$$Ui=(ui1, ui2, ..., uig)$$

Ｕｉ：文書の各分類グループへ所属度ベクトル
ｕｉｃ：文書ｉの分類グループｃへの所属度

数 3

The example of whenever [document classification affiliation] is shown in Table 2. Whenever [affiliation / of each document] (U1, U2, U3 --) is shown to each classification group (G1, G2, G3--) of every by this example.

[0066]

[Table 2]

	G1	G2 ... Gg
U1	0.7	0.2 ...
U2	0.8	0.1 ...
U3	0.3	0.6 ...
...		

文書分類所属度の例

表 2

Generation of classification theme information is performed by two kinds of methods as follows in generation processing (step 504) of the continuing classification theme information.

[0067] 1) The keyword method of keyword ****** is a method which expresses each classification group's theme with the combination of some keywords, and can consider two kinds of methods as follows for the extract of a keyword in that case. The 1st method extracts k words of an element with the high weight in the representation document vector Vc of an applicable classification in order, and those words are used for it as information showing the group's theme. The 2nd method elects r document vectors as the high order of whenever [affiliation] to the document set of an applicable classification, extracts k words in the r document vector set sequentially from what has the high number of appearance documents, and they are used for it as information showing the group's theme information.

[0068] 2) By the text method of text ******, extract the text which owns most those keywords by character string collating per sentence using the keyword theme information acquired by the keyword

method to the text (a title is included) of r head paragraphs of a document elected in order for the above-mentioned keyword method to generate theme information, and use the text sentence as the group's theme information.

[0069] Thus, it will be shown to a user in predetermined presentation sequence so that it may mention later, theme information, i.e., classification theme information, of each obtained group (an above-mentioned keyword group or an above-mentioned title sentence etc.). Here, if the goodness of fit to the retrieval type of Ri and a classification group is set to GRc for the goodness of fit to the retrieval question of the searched document i, several 4 relation will be materialized among both.

[0070]

[Equation 4]

$$GRc = \sum_{i}^{rc} Ri / rc$$

rc：グループcに対して所属度の高い順に選出された文書数

Ri：グループcに対して選出されたrc個の文書集合内のi文書の適合度

数 4

Here, how to ask for the number rc (c= 1, --, g;g: the number of classifications) of documents which was shown in several 4 and which was elected as the high order of whenever [affiliation] to Group c is shown in the flow chart of drawing 7 . If processing is started, after initializing rc in this drawing (step 701) (rc=0), Whenever [greatest affiliation] is called for from the line data Ui of whenever [affiliation / Of Document i] (step 702). The number rc of members of the maximum and the corresponding group c is added (step 703). The above processing (step 702,703) adding i every [1 / +], it will be repeated until the aggregate value serves as i=n (the number of documents) (step 705 YES), and as a result, finally, the value of rc will be calculated (step 704).

[0071] Thus, completion of generation (the decision of presentation sequence is included) of classification theme information starts dynamic display processing of the retrieval result using the theme classification information searched for (step 505). (step 504)

[0072] The detail of the dynamic display process of a retrieval result is shown in the flow chart of drawing 8 . If processing is started in this drawing, as the retrieval result output area A2 set up on the screen of the image display machine which constitutes retrieval equipment is shown in drawing 9 or drawing 10 , 2 ****s will be carried out up and down, and, thereby, the window W4 for theme classification information displays (Subject Window) and the window W5 for a retrieval result output (Hit Window) will appear. and the window W4 for theme classification information displays (Subject Window) -- setting -- predetermined display voice -- presentation of each classification theme information is performed more like (step 801). As mentioned above, presentation of each of this classification theme information is performed by a keyword method and the text method.

[0073] An example of the display screen by the keyword method is shown in drawing 9 R> 9. In addition, the searched text set is divided into five classification groups in this example. In the window W4 for theme classification information displays (Subject Window), as shown in this drawing, as the left brink section is met, five group carbon buttons 25-29 corresponding to the classification group number "1" - classification group number "5" are arranged at the vertical single tier, and the keyword groups 30-34 which express the theme of the classification group concerned exactly are arranged on the right-hand side of those group carbon buttons 25-29. In this example, on the right-hand side of the group carbon button 25 corresponding to the classification group number "1" As a keyword group 30, "SINGAPORE;CANE;PUNISH;US" is displayed and on the right-hand side of the group carbon button 26 corresponding to the classification group number "2" As a keyword group 31, "DALAILAMA;MEET;CHINA;TIBET" is displayed and on the right-hand side of the group carbon button

27 corresponding to the classification group number "3" As a keyword group 32, "MEET;LEADER;GOVERNMENT;OFFICIAL" is displayed and on the right-hand side of the group carbon button 28 corresponding to the classification group number "4" as a keyword group 33 "NIXON;NATION;SINGAPORE;DIRECTIVE" is displayed and on the right-hand side of the group carbon button 29 corresponding to the classification group number "5" "QUESTION;CHARACTER;PEOPLE;POLITICS" is displayed as a keyword group 34.

[0074] Moreover, such theme classification information is arranged sequentially from what has a high goodness of fit with a retrieval question (Query) according to the presentation sequence searched for previously. That is, the theme symbolized with the classification group number "1" in this example will have the highest goodness of fit with a retrieval question, and a goodness of fit with a retrieval question will be the lowest for the theme symbolized with the classification group number "5." Therefore, from the display sequence in the window W4 for theme classification information displays (Subject Window), a user 7 can know easily the classification group nearest to the information which he is looking for, and can check each classification group's theme roughly based on the contents of the keyword groups 30-34 which moreover express each contents directly. And as the original retrieval question is met, retrieval narrowing down can be efficiently performed by starting classification result display processing (step 802), so that it may explain in detail later.

[0075] An example of the display screen by the text method is shown in drawing 10 R> 0. In addition, the searched text set is divided into five classification groups also in this example. In the window W4 for theme classification information displays (Subject Window), as shown in this drawing, as the left brink section is met, five group carbon buttons 25-29 corresponding to the classification group number "1" - classification group number "5" are arranged at the vertical single tier, and the short text sentences 35-39 which express the theme of the classification group concerned exactly are arranged on the right-hand side of those group carbon buttons 25-29. In this example, on the right-hand side of the group carbon button 25 corresponding to the classification group number "1" "Clinton Protest Singapore Caning.Mulls Response" is displayed as a text sentence 35. On the right-hand side of the group carbon button 26 corresponding to the classification group number "2" As a text sentence 36, "Clinton Meets With Dalai Lama" is displayed and on the right-hand side of the group carbon button 27 corresponding to the classification group number "3" as a text sentence 37 "IndianLeader Meet Clinton" is displayed and on the right-hand side of the group carbon button 28 corresponding to the classification group number "4" As a text sentence 38, "Nixon Had LivingWill" is displayed and on the right-hand side of the group carbon button 29 corresponding to the classification group number "5" "Clinton News Conferens-Text" is displayed as a text sentence 39.

[0076] Moreover, according to the presentation sequence searched for previously, it is arranged also about such theme classification information sequentially from what has a high goodness of fit with a retrieval question (Query). That is, a classification group's theme symbolized with the classification group number "1" in this example will have the highest goodness of fit with a retrieval question, and a goodness of fit with a retrieval question will be the lowest for a classification group's theme symbolized with the classification group number "5." Therefore, from the display sequence in the window W4 for theme classification information displays (Subject Window), a user 7 can know easily the classification group nearest to the information which he is looking for, and can check each classification group's theme roughly based on the contents of the text sentences 35-39 which moreover express each contents directly. And as the original retrieval question is met, retrieval narrowing down can be efficiently performed by starting classification result display processing (step 802), so that it may explain in detail later.

[0077] Next, the processing for the last display of a retrieval result using Ui is explained to a detail whenever [to each classification group of each document obtained by the fuzzy clustering explained previously / affiliation]. In addition, in this example, three kinds of processings are prepared for the last display of a classification result, and these processings are started in the screen shown in drawing 9 or

drawing 10 by operating any one of the group carbon buttons 25-29 (step 802).
[0078] it explained previously -- as -- the description matrix of the document set as a retrieval result in this invention -- receiving -- FCM -- fuzzy clustering is performed using law and, thereby, Ui is calculated whenever [to each classification category of each document / affiliation]. Now, temporarily, five documents (001, 002, 003, 004, 005) exist, and it is assumed about each of those documents that it is as whenever [to each of three classification categories (a category 1, a category 2, category 3) / affiliation] being Table 3.
[0079]
[Table 3]

文書番号	カテゴリ1	カテゴリ2	カテゴリ3
001	0.50	0.30	0.20
002	0.60	0.10	0.30
003	0.10	0.80	0.10
004	0.25	0.34	0.41
005	0.35	0.10	0.55

割り付けの説明のための数値例

表3

3 kinds of display-processing (1) - (3) of a fuzzy classification result is explained to the origin of the above premise.
[0080] (1) If it is shown in this display processing when assigning to the classification category of k high orders of whenever [affiliation / of each document], each document (001-005) is assigned to k classification categories chosen sequentially from the high thing of whenever [affiliation]. If k= 1 (binary-ized method), about a document (001), in for example, the category 1 which is 0.50 whenever [maximum affiliation] About a document (002), in the category 1 which is 0.60 whenever [maximum affiliation] About a document (003), it is assigned [document / (004) / document / (005)] to the category 3 which is 0.55 whenever [maximum affiliation] by the category 3 which is 0.41 whenever [maximum affiliation] at the category 2 which is 0.80 whenever [maximum affiliation], respectively. this -- a classification category (G1, G2, G3) -- if it arranges independently -- category G1= (001 002) ;N1=2 category G2= (003) 2= ;N1 category G3= (004 005) ; -- several documents which serve as N 3= 2 and are contained in the classification group G1 -- several documents with which N1 is contained in two pieces and the classification group G2 -- several documents with which N2 are contained in one piece and the classification group 3 -- N3 are made into two pieces. And the document it was presupposed that it did in this way and was belonged to each category will be displayed in the window (HitWindow) W5 for a retrieval result output with assignment of the group number so that it may explain to a detail later.
[0081] An example of the program for realizing the above display processing (1) is shown in drawing 11 . If processing is started in this drawing, after performing setting processing (step 1101) of k value, and initialization processing (step 1102) of i, c, and Nc, The sorting application to the degree line data i of affiliation of Document i (step 1103), The processing which extracts the group number of k pieces sequentially from a data value whenever [maximum affiliation] (step 1104), And the processing (step 1105) which adds the number of members at the same time it registers Document i into k corresponding groups If it is repeated until a publication number i is set to n (step 1106), and a publication number i reaches n, the document allotment result for every group will be outputted, and processing will be completed (step 1107).
[0082] (2) If it is shown in this display processing when assigning to the classification category which has a value whenever [beyond a certain threshold alpha / affiliation], each document (001-005) is assigned to the classification category which has a value whenever [beyond a certain threshold alpha / affiliation]. Here, as alpha, it is possible to be referred to as 1/g (g: the number of classifications), for

example. In the example shown in Table 3, since it is set to g= 3 and alpha= 0.33, about a document (001), in the category 1 whose value is 0.33 or more whenever [affiliation] If attached to a document (002), by the same reason, for the reason same about a document (003), it is the reason same about a document (005), and is assigned to a category 1 by the category 2 in a category 1 and a category 3 at the reason same about a document (004) at a category 2 and a category 3. this -- a classification category (G1, G2, G3), if it arranges independently Category G1= (001, 002, 005) ; [N 1= 3] Category G2= (003 004) ; N 2= 2 Category G3= (004 005) several documents which serve as;N 3= 2 and are contained in the classification group G1 -- several documents with which N1 is contained in three pieces and the classification group G2 -- several documents with which N2 are contained in two pieces and the classification group 3 -- N3 are made into two pieces. And the document it was presupposed that it did in this way and was belonged to each category will be displayed in the window (HitWindow) W5 for a retrieval result output with assignment of the group number so that it may explain to a detail later.

[0083] An example of the program for realizing the above display processing (2) is shown in drawing 12 . If processing is started in this drawing, after performing setting processing (step 1201) of alpha value, and initialization processing (step 1202) of i, c, and Nc, The processing which extracts the group number of uic>alpha to the degree line data i of affiliation of Document i (step 1203), The processing (step 1204) which adds the number of members at the same time it registers Document i into each corresponding group If it is repeated until a publication number i is set to n (step 1205), and a publication number i reaches n, the document allotment result for every group will be outputted, and processing will be completed (step 1206).

[0084] (3) If it is shown in this display processing when assigning to a classification category in consideration of the probability distribution of a category, each document (001-005) is assigned to a classification category in consideration of the probability distribution of a category. Here, the probability distribution (Pc) of the classification category of a document is searched for according to several 5, and the number Nc of documents of Classification c is called for according to several 6.

[0085]

[Equation 5]

$$Pc = rc / n$$

rc : (1) の2値化方式により得られた分類cの文書数

n : 文書集合の全文書数

数 5

[Equation 6]

$$Nc = N(\alpha) \cdot Pc$$

$N(\alpha)$: (2) の割り付け方式により得られた各分類の文書表示数の和

数 6

In the example shown in Table 3, since it is set to P1=0.4, P2=0.2, and P3=0.4 and is set to N(0.33) = 7, it is set to N 1= 2.8 (about 3), N 2= 1.4 (about 1), and N 3= 2.8 (about 3). this -- a classification category (G1, G2, G3) -- if it arranges independently Category G1= (001, 002, 005) ;N 1= 2 Category G2= (003) ; N 2= 1 Category G3= (002, 004, 005) It is set to;N 3= 2. And the document it was presupposed that it did in this way and was belonged to each category will be displayed in the window W5 for a retrieval result output (Hit Window) with assignment of the group number so that it may explain to a detail later.

[0086] An example of the program for realizing the above display processing (3) is shown in drawing 13 . If processing is started in this drawing, setting processing (step 1301) of alpha value, initialization

processing (step 1302) of i, c, and Nc, processing (step 1303) that searches for the probability distribution (Pc=rc/n) of the classification category of a document, and processing (step 1304) which calculates Nc of the number of documents of Classification c will be performed one by one. Then, sorting application [as opposed to / whenever / affiliation / of Document c / string data uic] (step 1305), The processing which extracts the publication number of the member of Nc individual of correspondence in order from a value whenever [maximum affiliation] (step 1306), And the processing (step 1307) which registers the document of Nc individual into the group c of relevance If it is repeated until Classification c turns into a several g classification (step 1308 NO), and Classification c reaches a several g classification (step 1308 YES), the document allotment result for every group will be outputted, and processing will be completed (step 1309).

[0087] Next, the document assigned to each classification group in three kinds of either of allotment processing (1) – (3) which was explained above explains in what kind of mode it is displayed in the window W5 for a retrieval result output on the display screen (Hit Window).

[0088] It sets on the screen shown in drawing 9 , and they are one of group carbon buttons (in this example). If assignment actuation of the group carbon button 26 is carried out, the short text sentences (head part of the text concerned which contains a title etc. in this example) 40–44 equivalent to the document assigned to each classification group in three kinds of either of allotment processing (1) – (3) which was mentioned above It will be displayed in the window W5 for a retrieval result output (Hit Window) (step 802).

[0089] Namely, by having specified the classification group number "2" symbolized by the keyword group 31 ("DALAILAMA;MEET;CHINA;TIBET") in this example In the window W5 for a retrieval result output (Hit Window) Five text sentences 40 ("Clinton Meets With Dalai Lama) relevant to this, The text sentence 41 ("Clinton, Gore MeetDalai Lama on Tibetan Right), The text sentence 42 ("China Warns Clinton NottoMeet Dalai Lama"), The text sentence 43 ("Clinton May Meet Dalai Lama before China Decision") and the text sentence 44 ("Indian Leader Meet Clinton") are displayed. And since the order assignment carbon button 51 of a group goodness of fit described as "G" among drawing is operated, these text sentences 40–44 are arranged in order of a goodness of fit with the classification group symbolized with the specified classification group number "2" concerned, and are displayed. In addition, the scroll bar of the windows W4 and W5 where signs 45 and 46 are located in the left-hand side, respectively, and 49 are the displays of classification group number.

[0090] [in the window W5 for a retrieval result output (Hit Window)] furthermore, into each head part of each text sentences 40–44 Three kinds of goodness of fit marks which express the goodness of fit which each text sentences 40–44 have to the classification group concerned to a three-stage (47a, 47b, 47c), Three kinds of goodness of fit marks (48a, 48b, 48c) which express the goodness of fit which each text sentences 40–44 have to the retrieval question 22 concerned to a three-stage are displayed. The shape of a basic form of the goodness of fit mark (47a, 47b, 47c) which expresses a goodness of fit with the classification group concerned with this example is SNOW BRAND. About goodness of fit mark 47a equivalent to a goodness of fit "quantity", the small-circle form part of the core to black painting out About goodness of fit mark 47c which is further equivalent to a goodness of fit "low" in the small-circle form part of the core at gray painting out about goodness of fit mark 47b equivalent to a goodness of fit "inside", the small-circle form part of the core is considered as void. Moreover, the shape of a basic form is a round mark, and the goodness of fit mark (48a, 48b, 48c) showing a goodness of fit with the retrieval question concerned is taken as void about goodness of fit mark 48c which is further equivalent to a goodness of fit "low" at gray painting out about goodness of fit mark 48b which is equivalent to a goodness of fit "inside" at black painting out about goodness of fit mark 48a equivalent to a goodness of fit "quantity."

[0091] Therefore, a user 7 can find out the target information exactly according to the contents 40–44 of a display in this window W5 for a retrieval result output (Hit Window), checking the text set belonging to the group of the classification group number "2" in the text set which it is as a result of retrieval

sequentially from what has a high goodness of fit with this classification group "2" by considering a goodness of fit mark (47a, 47b, 47c) as reliance. In addition, since the goodness of fit of each text sentences 40-44 and the retrieval question 22 can also be known by referring to a goodness of fit mark (48a, 48b, 48c), much more positive retrieval narrowing down can be performed by considering both marks 47 and 48 as reference. In addition, although not illustrated, when the order assignment carbon button 50 of a retrieval question goodness of fit described as "R" among drawing is operated, in drawing 8 , classification subject guide assignment processing (step 804) is performed, and each text sentences 40-44 will be arranged in order of a goodness of fit with the retrieval question 22 concerned, and will be displayed. Therefore, a retrieval result can be checked along the desired retrieval direction by any of the order assignment carbon button 50 of a retrieval question goodness of fit, and the order assignment carbon button 51 of a group goodness of fit are chosen, changing the array of each text sentences 40-44.

[0092] In the condition that the retrieval result shown in drawing 9 is displayed on the other hand, if the actuation exchange demand carbon button (HELP) 15 is operated, it will return to drawing 8 and subject guide option processing (step 805) will be performed, and the display in the window W4 for theme classification information displays (Subject Window) will change from the above-mentioned keyword method to a text method, as shown in drawing 10 . Therefore, by the keyword method, even when it is hard to grasp the contents of the classification group concerned, according to presenting of the theme classification information by this text method, the theme symbolized into the classification group concerned can be known more exactly. In addition, when an indicative data is not settled in each window W4 and W5, it cannot be overemphasized that it can check scrolling the contents of a display by actuation of scroll bars 45 and 46.

[0093]

[Effect of the Invention] By the above explanation, it can become also to the guide of how by the ability being able to aim at improvement in the retrieval precision by prevention of the leakage in retrieval in improvement in retrieval effectiveness, and a list by make the check over a document retrieval result easy according to this invention, the theme information moreover showed extracts data efficiently, and puts them so that clearly, and advanced adaptation retrieval (Relevance Feedback) can be make to perform using this added response indication.

[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1] It is the block diagram showing notionally the configuration of the text browsing equipment with which this invention approach and equipment were applied.

[Drawing 2] It is the explanatory view showing notionally an operation of the fuzzy grouping section (Fuzzy Clustering) and the theme classification information generation section (Subject Generation).

[Drawing 3] It is the General flow chart which shows roughly the whole actuation of the text browsing equipment concerning this invention.
[Drawing 4] It is the screen explanatory view showing the condition of having performed retrieval actuation without grouping processing in the text browsing equipment concerning this invention.
[Drawing 5] It is the flow chart shown focusing on generation processing of the theme classification information in the text browsing equipment concerning this invention.
[Drawing 6] It is the explanatory view showing notionally generation of the document abstraction in the text browsing equipment concerning this invention, and a document vector.
[Drawing 7] It is the flow chart which shows the processing for asking for Group's c number rc of members in the text browsing equipment concerning this invention.
[Drawing 8] It is the flow chart which shows dynamic processing of the retrieval result using the theme classification information in the text browsing equipment concerning this invention.
[Drawing 9] It is the screen explanatory view showing the condition of having performed retrieval actuation accompanied by grouping processing by the keyword method in the text browsing equipment concerning this invention.
[Drawing 10] It is the screen explanatory view showing the condition of having performed retrieval actuation accompanied by grouping processing by the text method in the text browsing equipment concerning this invention.
[Drawing 11] It is the flow chart which sets with the text browsing equipment concerning this invention to display a retrieval result according to a group, and shows the allotment processing to the classification category of k high orders of whenever [affiliation / of each document].
[Drawing 12] It is the flow chart which shows the allotment processing to the classification category which sets with the text browsing equipment concerning this invention to display a retrieval result according to a group, and has a value whenever [beyond alpha value / affiliation].
[Drawing 13] the rate to the classification category which set with the text browsing equipment concerning this invention to display a retrieval result according to a group, and took the probability distribution of a category into consideration -- market-making -- it is the flow chart which shows processing.
[Description of Notations]
1 Retrieval Question Input Section
2 Document Database
3 Search Engine
4 Extracted Related Text Set
4a, 4b, 4c Classification group
5 Processing Processing Section
5a Fuzzy grouping section
5b Theme classification information generation section
6 Retrieval Result Presentation Section
7 User
12 Input Guide Sentence
13 Start Button
14 Cancellation Carbon Button
15 Help Button
16 Scroll Bar
17 Whole Sentence Demand Carbon Button
18 The QBE Carbon Button
19 Classification-ized Demand Carbon Button
20 Abstract Demand Carbon Button
21 Reset Button

22 Retrieval Question
23 Title of Each Text Which Constitutes Text Set
24a, 24b, 24c Goodness of fit mark
25–29 Group carbon button
30–34 Keyword group
35–39 Text sentence
40–44 Text sentence
45 46 Scroll bar
49 Display of Classification Group Number
47a, 47b, 47c Goodness of fit mark for every group
48a, 48b, 48c Goodness of fit mark to a retrieval question
49 Display of Classification Group Number
50 The Order Assignment Carbon Button of Retrieval Question Goodness of Fit
51 The Order Assignment Carbon Button of Group Goodness of Fit
A Original retrieval question
B Each extracted configuration text
calcium, Cb, Cc Theme classification information
Da, Db, Dc A group's center of gravity
A1 Retrieval question input area
A2 Retrieval result output area
W2 Window for a retrieval question input
W3 Window for a retrieval result output
W4 Window for theme classification information displays
W5 Window for a retrieval result output

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号

特開平9－231238

(43)公開日　平成９年(1997)９月５日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30			Ｇ０６Ｆ	15/403	３７０Ａ
				15/401	３１０Ｄ
				15/403	３８０Ｅ

審査請求　未請求　請求項の数24　ＦＤ　（全 19 頁）

(21)出願番号　特願平8－58391

(22)出願日　平成８年(1996)２月20日

(71)出願人　000002945
オムロン株式会社
京都府京都市右京区花園土堂町10番地

(72)発明者　呉　亜棟
京都府京都市右京区花園土堂町10番地　オムロン株式会社内

(72)発明者　十河　太治
京都府京都市右京区花園土堂町10番地　オムロン株式会社内

(72)発明者　澤田　晃
京都府京都市右京区花園土堂町10番地　オムロン株式会社内

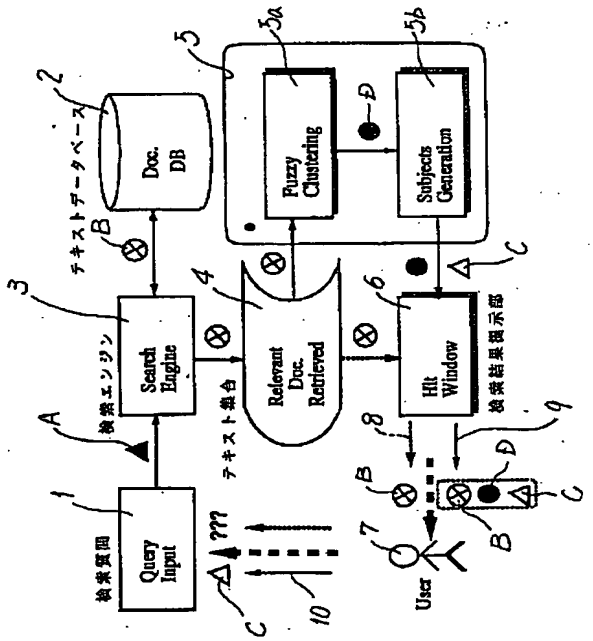
(74)代理人　弁理士　飯塚　信市

(54)【発明の名称】　テキスト検索結果表示方法及び装置

(57)【要約】

【課題】　文書検索結果に対する確認を容易として、検索効率の向上、並びに、検索漏れの防止による検索精度の向上を図ることができ、しかも、提示された主題情報がデータを如何に効率的に絞り込めるかの指針にもなり、この付加された応答情報を利用して高度な適応検索（Relevance Feedback）を行い得る。

【解決手段】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割し、該分割により得られた各グループのそれぞれについて、当該グループの属性を表現する主題分類情報を生成し、該生成された各グループの主題分類情報をグループ別に区分して表示する。



【特許請求の範囲】

【請求項１】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップと、

前記分割ステップによって得られた各グループのそれぞれについて、当該グループの属性を表現する主題分類情報を生成する生成ステップと、

前記生成ステップで求めた各グループの主題分類情報をグループ別に区分して表示する表示ステップとを具備する、

ことを特徴とするテキスト検索結果表示方法。

【請求項２】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップと、

前記分割ステップによって得られた各グループのそれぞれについて、当該グループの属性を表現する主題分類情報を生成する生成ステップと、

前記各グループのそれぞれについて、そのグループと前記検索条件との間の適合度を求めるグループ適合度算出ステップと、

前記生成ステップで求めた各グループの主題分析情報を、前記適合度算出ステップによって求めた適合度の大きい順に、グループ別に区分して表示する表示ステップとを具備する、

ことを特徴とするテキスト検索結果表示方法。

【請求項３】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップと、

前記グループ内の各テキストの内容の分析結果に基いて、各テキストの当該グループに対する所属度を算出する所属度算出ステップと、

前記複数個のグループの中で、テキスト表示対象となるグループを選択するための選択ステップと、

前記選択ステップで選択されたグループ内のテキストを前記算出された所属度の順に内容表示する表示ステップとを具備する、

ことを特徴とするテキスト検索結果表示方法。

【請求項４】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップと、

前記グループ内の各テキストの内容の分析結果に基いて、各テキストの前記検索条件に対する適合度を算出する適合度算出ステップと、

前記複数個のグループの中で、テキスト表示対象となるグループを選択するための選択ステップと、

前記選択ステップで選択されたグループ内のテキストを

前記算出された適合度の順に内容表示する表示ステップとを具備する、

ことを特徴とするテキスト検索結果表示方法。

【請求項５】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップと、

前記グループ内の各テキストの内容の分析結果に基いて、各テキストの当該グループに対する所属度を算出する所属度算出ステップと、

前記グループ内の各テキストの内容の分析結果に基いて、各テキストの前記検索条件に対する適合度を算出する適合度算出ステップと、

前記複数個のグループの中で、テキスト表示対象となるグループを選択するための表示対象グループ選択ステップと、

前記各グループ内のテキストを検索条件への適合度順に表示するか、或いは当該グループへの所属度の順に表示するかを選択するための表示順序基準選択手段と、

前記表示対象グループ選択ステップで選択されたグループ内のテキストを前記表示順序基準選択手段にて選択された表示順序基準の順に内容表示する表示ステップとを具備する、

ことを特徴とするテキスト検索結果表示方法。

【請求項６】　前記分割ステップは、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を、ファジイ・クラスタリング法を用いて複数個のグループに分割する、

ことを特徴とする請求項１乃至請求項５のいずれかに記載のテキスト検索結果表示方法。

【請求項７】　前記生成ステップにて生成される当該グループの属性を表現する主題分類情報は、当該グループの属性を幾つかのキーワードの組により表すものである、

ことを特徴とする請求項１若しくは請求項２のいずれかに記載のテキスト検索結果表示方法。

【請求項８】　前記生成ステップにて生成される当該グループの属性を表現する主題分類情報は、当該部ループの属性を短い文章により表すものである、

ことを特徴とする請求項１若しくは請求項２のいずれかに記載のテキスト検索結果表示方法。

【請求項９】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合の特徴行列に対してファジイ・クラスタリングを行い、各文書毎に各分類カテゴリーへの所属度を生成する所属度生成ステップと、

前記生成された所属度を用いて、各文書を１若しくは２以上の分類カテゴリーに割り付ける文書割り付けステップと、

前記複数個の分類カテゴリーの中で、テキスト表示対象

となる分類カテゴリーを選択するための分類カテゴリー選択ステップと、
前記分類カテゴリー選択ステップで選択された分類カテゴリー内のテキストをそのグループに対する適合度の順に内容表示する表示ステップとを具備する、
ことを特徴とするテキスト検索結果表示方法。
【請求項１０】　前記文書割り付けステップは、各文書をその所属度の上位ｋ個の分類カテゴリーに割り付ける、
ことを特徴とする請求項９に記載のテキスト検索結果表示方法。
【請求項１１】　前記文書割り付けステップは、各文書をある閾値α以上の所属度値を有する分類カテゴリーに割り付ける、
ことを特徴とする請求項９に記載のテキスト検索結果表示方法。
【請求項１２】　前記文書割り付けステップは、各文書をカテゴリーの確率分布を考慮して分類カテゴリーに割り付ける、
ことを特徴とする請求項９に記載のテキスト検索結果表示方法。
【請求項１３】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割手段と、
前記分割手段によって得られた各グループのそれぞれについて、当該グループの属性を表現する主題分類情報を生成する生成手段と、
前記生成手段で求めた各グループの主題分類情報をグループ別に区分して表示する表示手段とを具備する、
ことを特徴とするテキスト検索結果表示装置。
【請求項１４】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割手段と、
前記分割手段によって得られた各グループのそれぞれについて、当該グループの属性を表現する主題分類情報を生成する生成手段と、
前記各グループのそれぞれについて、そのグループと前記検索条件との間の適合度を求めるグループ適合度算出手段と、
前記生成手段で求めた各グループの主題分析情報を、前記適合度算出手段によって求めた適合度の大きい順に、グループ別に区分して表示する表示手段とを具備する、
ことを特徴とするテキスト検索結果表示装置。
【請求項１５】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割手段と、
前記グループ内の各テキストの内容の分析結果に基いて、各テキストの当該グループに対する所属度を算出する所属度算出手段と、
前記複数個のグループの中で、テキスト表示対象となるグループを選択するための選択手段と、
前記選択手段で選択されたグループ内のテキストを前記算出された所属度の順に内容表示する表示手段とを具備する、
ことを特徴とするテキスト検索結果表示装置。
【請求項１６】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割手段と、
前記グループ内の各テキストの内容の分析結果に基いて、各テキストの前記検索条件に対する適合度を算出する適合度算出手段と、
前記複数個のグループの中で、テキスト表示対象となるグループを選択するための選択手段と、
前記選択手段で選択されたグループ内のテキストを前記算出された適合度の順に内容表示する表示手段とを具備する、
ことを特徴とするテキスト検索結果表示装置。
【請求項１７】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割手段と、
前記グループ内の各テキストの内容の分析結果に基いて、各テキストの当該グループに対する所属度を算出する所属度算出手段と、
前記グループ内の各テキストの内容の分析結果に基いて、各テキストの前記検索条件に対する適合度を算出する適合度算出手段と、
前記複数個のグループの中で、テキスト表示対象となるグループを選択するための表示対象グループ選択手段と、
前記各グループ内のテキストを検索条件への適合度順に表示するか、或いは当該グループへの所属度の順に表示するかを選択するための表示順序基準選択手段と、
前記表示対象グループ選択手段で選択されたグループ内のテキストを前記表示順序基準選択手段にて選択された表示順序基準の順に内容表示する表示手段とを具備する、
ことを特徴とするテキスト検索結果表示装置。
【請求項１８】　前記分割手段は、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を、ファジイ・クラスタリング法を用いて複数個のグループに分割する、
ことを特徴とする請求項１３乃至請求項１７のいずれかに記載のテキスト検索結果表示装置。
【請求項１９】　前記生成手段にて生成される当該グループの属性を表現する主題分類情報は、当該グループの

属性を幾つかのキーワードの組により表すものである、
ことを特徴とする請求項１３若しくは請求項１４のいずれかに記載のテキスト検索結果表示装置。

【請求項２０】　前記生成手段にて生成される当該グループの属性を表現する主題分類情報は、当該グループの属性を短い文章により表すものである、
ことを特徴とする請求項１３若しくは請求項１４のいずれかに記載のテキスト検索結果表示装置。

【請求項２１】　与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合の特徴行列に対してファジイ・クラスタリングを行い、各文書毎に各分類カテゴリーへの所属度を生成する所属度生成手段と、
前記生成された所属度を用いて、各文書を１若しくは２以上の分類カテゴリーに割り付ける文書割り付け手段と、
前記複数個の分類カテゴリーの中で、テキスト表示対象となる分類カテゴリーを選択するための分類カテゴリー選択手段と、
前記分類カテゴリー選択手段で選択された分類カテゴリー内のテキストをそのグループに対する適合度の順に内容表示する表示手段とを具備する、
ことを特徴とするテキスト検索結果表示装置。

【請求項２２】　前記文書割り付け手段は、各文書をその所属度の上位ｋ個の分類カテゴリーに割り付ける、
ことを特徴とする請求項２１に記載のテキスト検索結果表示装置。

【請求項２３】　前記文書割り付け手段は、各文書をある閾値α以上の所属度値を有する分類カテゴリーに割り付ける、
ことを特徴とする請求項２１に記載のテキスト検索結果表示装置。

【請求項２４】　前記文書割り付け手段は、各文書をカテゴリーの確率分布を考慮して分類カテゴリーに割り付ける、
ことを特徴とする請求項２１に記載のテキスト検索結果表示装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、文書データベースの検索に好適なテキスト検索結果表示方法及び装置に係り、特に、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割し、該分割により得られた各グループの属性を表現する主題分類情報をグループ別に区分して表示するようにしたテキスト検索結果表示方法及び装置に関する。

【０００２】

【従来の技術】従来のテキスト検索結果表示装置としては、例えば、特開平６－７６００４号公報に記載されたものが知られている。

【０００３】この装置は、データベース検索結果を格納するデータベース検索解格納部と、前記データベース検索解が有する複数の属性値に利用者の制御入力を加味して各検索解間の距離を算出する検索解間距離算出手段と、検索解間距離を用いて検索解を利用者に指定した個数或いは予め定められた個数のグループに分割する検索解グループ分割手段と、所属グループの重心付近に位置する検索解を算出するグループ代表検索解算出手段と、各グループの代表検索解の中から特定の検索解を利用者に選択させる代表検索解選択手段と、代表検索解が属しているグループ内の全検索解を表示するグループ内検索解表示手段とから構成されている。

【０００４】すなわち、この従来装置にあっては、non-overlapping手法で構造化された（数値）データベース検索解をユーザーの指定した分類数に分類するもので、分類されたグループの重心に最も近い検索解を１件ずつパイロットデータとして表示して利用者に希望するグループを選択させ（順位付けなし）、選択されたグループ内の全検索解をランキングせずに表示するものである。

【０００５】

【発明が解決しようとする課題】しかしながら、このような従来のテキスト検索結果表示装置（検索解表示装置）にあっては、次の理由により、フルテキストのような非構造化データベースへの適用は困難であると言う問題点があった。

【０００６】すなわち、このような従来装置にあっては、グループ内の重心位置の代表検索解が表示されるため、代表検索解がグループ内の代表文書である場合には、その代表文書の内容を端的に表すものが表示されず、文書全体が表示されるのでグループの内容が把握し難い。つまり、分類された各グループの主題意味を提示するために、単なるグループの重心に最も近い検索解を１件ずつパイロットデータとして表示するだけでは、内容的に特定すぎる場合があり、むしろ、グループ内の共通的な属性項目群を抽出し、利用者に提示することが好ましい。加えて、フルテキスト検索システムの場合にあっては、パイロットデータとして全ての属性データをそのまま提示することは無意味であり、文書内容を容易に理解できるようなパイロットデータの新しい定義が望まれる。

【０００７】また、従来装置にあっては、グループが検索条件に対する適合度の順に並べられないので、検索目的に合致したグループを選択し難い。加えて、従来装置にあっては、グループ内の解がグループへの所属度の順に並んでいないので、グループの代表解を参照するだけでは、グループのイメージが把握し難い場合でも、他の解を参照してイメージを把握することが困難である。つまり、選択されたグループ内の全検索解をランキングせずに表示する方式では、分類件数が多くなると、検索結

果への特定のために利用者の負担が大きくなる。このような負担を軽減して検索効率を向上させるためには、検索結果への特定を促進できるようなランキング機能が望まれる。

【０００８】更に、文書は複数の主題を持っているのが通例であるため、一つの文書を一つのクラスタにしか分類できない従来の手法では、文書分類結果の表示上では検索結果に漏れを生じる虞れがある。そのため、文書検索結果集合に対し主題分類を行う際に複数の異なる（主題を表す）クラスタに属することを許すようなoverlapping手法が望まれる。

【０００９】この発明は、上述の問題点に鑑みてなされたものであり、その目的とするところは、文書検索結果に対する確認を容易として、検索効率の向上、並びに、検索漏れの防止による検索精度の向上を図ることができ、しかも、提示された主題情報がデータを如何に効率的に絞り込めるかの指針にもなり、この付加された応答情報を利用して高度な適応検索（Relevance Feedback）を行い得るようにした検索結果表示方法及び装置を提供することにある。

【００１０】

【課題を解決するための手段】この出願の請求項１（又は請求項１３）に記載の発明は、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップ（又は手段）と、前記分割ステップ（又は手段）によって得られた各グループのそれぞれについて、当該グループの属性を表現する主題分類情報を生成する生成ステップ（又は手段）と、前記生成ステップ（又は手段）で求めた各グループの主題分類情報をグループ別に区分して表示する表示ステップ（又は手段）とを具備する、ことを特徴とするテキスト検索結果表示方法（又は装置）にある。

【００１１】ここで、『データベース』とは、ハードディスクや光ディスク等の大容量記憶媒体に記憶されたテキスト集合やインターネット上に存在するホームページ等のテキスト集合がこれに相当する。

【００１２】また、『主題分析』とは、テキストの内容を端的に示す情報を生成することを意味するものであり、文書内のタイトル上のキーワードの集合を生成するものであっても良い。実施の形態においては、文書を文書空間での特徴ベクトルで表現しているベクトル（Ｆｉ）がこれに相当する。

【００１３】また、『主題分類情報』とは、テキストのグループについて、そのグループの内容を端的に示す情報を意味する。実施の形態では、キーワード方式とテキスト方式との２方式が示されている。

【００１４】そして、この請求項１（又は請求項１３）の発明によれば、グループを端的に表現する情報を付加してグループ別に区分表示するので、検索結果を構成するグループの全体像を把握し易くなり、次の処理のためのグループ選択が非常に容易となる。

【００１５】この出願の請求項２（又は請求項１４）の発明は、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップ（又は手段）と、前記分割ステップ（又は手段）によって得られた各グループのそれぞれについて、当該グループの属性を表現する主題分類情報を生成する生成ステップ（又は手段）と、前記各グループのそれぞれについて、そのグループと前記検索条件との間の適合度を求めるグループ適合度算出ステップ（又は手段）と、前記生成ステップ（又は手段）で求めた各グループの主題分析情報を、前記適合度算出ステップによって求めた適合度の大きい順に、グループ別に区分して表示する表示ステップ（又は手段）とを具備する、ことを特徴とするテキスト検索結果表示方法（又は装置）にある。

【００１６】そして、この請求項２（又は請求項１４）の発明によれば、前記請求項１（又は請求項１３）に記載の発明の効果に加えて、検索条件への適合度の順に表示するので、検索目的に合致したグループをグループの内容を確認しながら選択することができる。

【００１７】この出願の請求項３（又は請求項１５）の発明は、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップ（又は手段）と、前記グループ内の各テキストの内容の分析結果に基いて、各テキストの当該グループに対する所属度を算出する所属度算出ステップ（又は手段）と、前記複数個のグループの中で、テキスト表示対象となるグループを選択するための選択ステップ（又は手段）と、前記選択ステップ（又は手段）で選択されたグループ内のテキストを前記算出された所属度の順に内容表示する表示ステップ（又は手段）とを具備する、ことを特徴とするテキスト検索結果表示方法（又は装置）にある。

【００１８】そして、この請求項３（又は請求項１５）の発明によれば、選択されたグループ内のテキストがグループへの所属度の順に表示されるので、グループの定義が把握し易くなる。

【００１９】この出願の請求項４（又は請求項１６）の発明は、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップ（又は手段）と、前記グループ内の各テキストの内容の分析結果に基いて、各テキストの前記検索条件に対する適合度を算出する適合度算出ステップ（又は手段）と、前記複数個のグループの中で、テキスト表示対象となるグループを選択するための選択ステッ

プ（又は手段）と、前記選択ステップ（又は手段）で選択されたグループ内のテキストを前記算出された適合度の順に内容表示する表示ステップ（又は手段）とを具備する、ことを特徴とするテキスト検索結果表示方法（又は装置）にある。

【００２０】そして、この請求項４（又は請求項１６）の発明によれば、検索条件に適したグループを選択し、さらにその中のテキストを検索条件の順に表示するので、検索結果をグループ分けしないでテキストを適合度順に表示する場合よりも、検索条件に対して適切なテキストが早く確実に表示される。

【００２１】この出願の請求項５（又は請求項１７）の発明は、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割ステップ（又は手段）と、前記グループ内の各テキストの内容の分析結果に基いて、各テキストの当該グループに対する所属度を算出する所属度算出ステップ（又は手段）と、前記グループ内の各テキストの内容の分析結果に基いて、各テキストの前記検索条件に対する適合度を算出する適合度算出ステップ（又は手段）と、前記複数個のグループの中で、テキスト表示対象となるグループを選択するための表示対象グループ選択ステップ（又は手段）と、前記各グループ内のテキストを検索条件への適合度順に表示するか、或いは当該グループへの所属度の順に表示するかを選択するための表示順序基準選択ステップ（又は手段）と、前記表示対象グループ選択ステップで選択されたグループ内のテキストを前記表示順序基準選択手段にて選択された表示順序基準の順に内容表示する表示ステップ（又は手段）とを具備する、ことを特徴とするテキスト検索結果表示方法（又は装置）にある。

【００２２】そして、この請求項５（又は請求項１７）の発明によれば、ユーザーの目的に応じてテキストの表示順序を変えることができる。

【００２３】この出願の請求項６（又は請求項１８）に記載の発明は、請求項１（又は請求項１３）乃至請求項５（又は請求項１７）のいずれかに記載のテキスト検索結果表示方法（又は装置）において、前記前記分割ステップ（又は手段）は、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合を、ファジイ・クラスタリング法を用いて複数個のグループに分割する、ことを特徴とするものである。

【００２４】そして、この請求項６（又は請求項１８）に記載の発明によれば、ある検索式により探し出された文書集合に対して自動的にorverlapping方式で主題内容によるファジイ分類（主題分類）が行われるため、検索漏れの防止による検索精度の向上が期待できる。

【００２５】この出願の請求項７（又は請求項１９）に記載の発明は、請求項１（又は請求項１３）又は請求項２（又は請求項１４）に記載のテキスト検索結果表示方法（又は装置）において、前記生成ステップ（又は手段）にて生成される当該グループの属性を表現する主題分類情報は、当該グループの属性を幾つかのキーワードの組により表すものである、ことを特徴とするものである。

【００２６】そして、この請求項７（又は請求項１９）に記載の発明によれば、当該グループの属性を幾つかのキーワードの組を通して直観的に把握することができる。

【００２７】この出願の請求項８（又は請求項２０）に記載の発明は、請求項１（又は請求項１３）又は請求項２（又は請求項１４）に記載のテキスト検索結果表示方法（又は装置）において、前記生成ステップ（又は手段）にて生成される当該グループの属性を表現する主題分類情報は、当該グループの属性を短い文章により表すものであることを特徴とするものである。

【００２８】そして、この請求項８（又は請求項２０）に記載の発明によれば、当該グループの属性を短い文章を通して誰にでも判りやすく理解させることができる。

【００２９】この出願の請求項９（又は請求項２１）に記載の発明は、与えられた検索条件に基いてデータベースを検索することにより得られたテキスト集合の特徴行列に対してファジイ・クラスタリングを行い、各文書毎に各分類カテゴリーへの所属度を生成する所属度生成ステップ（又は手段）と、前記生成された所属度を用いて、各文書を１若しくは２以上の分類カテゴリーに割り付ける文書割り付けステップ（又は手段）と、前記複数個の分類カテゴリーの中で、テキスト表示対象となる分類カテゴリーを選択するための分類カテゴリー選択ステップ（又は手段）と、前記分類カテゴリー選択ステップ（又は手段）で選択された分類カテゴリー内のテキストをそのグループに対する適合度の順に内容表示する表示ステップ（又は手段）とを具備する、ことを特徴とするテキスト検索結果表示方法（又は装置）にある。

【００３０】そして、この請求項９（又は請求項２１）に記載の発明によれば、overlapping手法を用いて各文書を１若しくは２以上の分類カテゴリーに割り付け、その状態にて選択された分類カテゴリー内のテキストをそのグループに対する適合度の順に内容表示するため、検索効率の向上、並びに、検索漏れの防止による検索精度の向上を図ることができる。

【００３１】この出願の請求項１０（又は請求項２２）に記載の発明は、前記請求項９（又は請求項２１）に記載の発明において、前記文書割り付けステップ（又は手段）は、各文書をその所属度の上位ｋ個の分類カテゴリーに割り付ける、ことを特徴とするものである。

【００３２】そして、この請求項１０（又は請求項２２）に記載の発明によれば、請求項９（又は請求項２１）に記載の発明の効果に加え、各分類カテゴリーにつ

いて常に所属度の高い順に一定個数の文書を表示させることができる。

【００３３】この出願の請求項１１（又は請求項２３）に記載の発明は、前記請求項９（又は請求項２１）に記載の発明において、前記文書割り付けステップは、各文書をある閾値α以上の所属度値を有する分類カテゴリーに割り付ける、ことを特徴とするものである。

【００３４】そして、この請求項１１（又は請求項２３）に記載の発明によれば、請求項９（又は請求項２１）に記載の発明の効果に加え、各分類カテゴリーについて常にある閾値α以上の所属度値を有する文書を表示させることができる。

【００３５】この出願の請求項１２（又は請求項２４）に記載の発明は、前記請求項９（又は請求項２１）に記載の発明において、前記文書割り付けステップは、各文書をカテゴリーの確率分布を考慮して分類カテゴリーに割り付ける、ことを特徴とするものである。

【００３６】そして、この請求項１２（又は請求項２４）に記載の発明によれば、請求項９（又は請求項２１）に記載の発明の効果に加え、各分類カテゴリーについてカテゴリーの確率分布を考慮して関連する文書を表示させることができる。

【００３７】

【発明の実施の形態】以下に、本発明方法及び装置の好適な実施の形態を添付図面を参照しながら詳細に説明する。

【００３８】先ず、本発明方法及び装置が適用されたテキスト検索装置の構成を図１の機能ブロック図により概念的に示す。同図において、１は検索開始時に入力されるべきオリジナル検索質問（Original Query）や検索絞り込み時に入力されるべきフィードバック検索質問（FeedBack Query）を入力するための検索質問入力部（Query Inputと記す）であり、具体的には、周知のように、マウスやキーボード等の操作部とそれらの信号を処理する入力用ソフトウェアにより構成される。

【００３９】２は検索対象となるテキスト集合に相当するテキスト（文書）データベース（Doc.DBと記す）であり、具体的には、ハードディスクや光ディスク等の大容量記憶媒体に記憶されたテキスト集合やインターネット上に存在するホームページ等のテキスト集合がこれに相当する。

【００４０】３はテキスト検索システムの中枢に位置する検索エンジン（Search Engineと記す）であり、具体的には、周知のように、前述の検索質問入力部１から入力されるオリジナル検索質問（Original Query）やフィードバック検索質問（FeedBack Query）を検索条件として所定のアルゴリズムに従って検索式を展開し、前述の文書データベース２から関連するテキスト集合を抽出するソフトウェアがこれに相当する。

【００４１】４はこのようにして検索エンジン（Search Engine）３により抽出された関連するテキスト集合（Relevant Doc. Retrievedと記す）であり、後述するように、このテキスト集合４が本発明における加工処理の対象となる。

【００４２】５は本発明の要部に相当する加工処理部であり、この加工処理部５はテキスト集合４を各テキストの主題分析結果を用いて自動的に複数個のグループに分割する分割手段に相当するファジイ集合化部（Fuzzy Clusteringと記す）５ａと、こうして得られた各グループのそれぞれについて、当該グループの属性を表現する主題分類情報を生成する主題分類情報生成部（Subject Generationと記す）５ｂとを中心として構成されている。

【００４３】ファジイ集合化部（Fuzzy Clustering）５ａ及び主題分類情報生成部（Subject Generation）５ｂの作用を図２に概念的に示す。同図において、符号４で示される実線にて囲まれた領域は検索エンジン（Search Engine）３にて抽出されたテキスト集合（Relevant Doc. Retrieved）の全体を表す。

【００４４】同様にして、符号４ａ，４ｂ，４ｃで示される破線にて囲まれた３つの領域はファジイ集合化部（Fuzzy Clustering）５にて分割された３つのグループのそれぞれを表す。

【００４５】符号Ａで示される黒塗り三角印は、検索開始時に入力されるオリジナル検索質問（Original Query）を表す。符号Ｂで示される×入り丸印は、オリジナル検索質問（Original Query）Ａの入力により検索抽出されたテキスト集合４の各構成テキストのそれぞれを表す。

【００４６】符号Ｃａ，Ｃｂ，Ｃｃで示される３個の白抜き三角印は、グループ４ａ，４ｂ，４ｃの属性を表現する主題分類情報（Group Subject）を表す。尚、これらの主題分類情報Ｃａ，Ｃｂ，Ｃｃは検索絞込みのために用いれば、フィードバック検索質問（FeedBack Query）としても好適なものである。

【００４７】符号Ｄａ，Ｄｂ，Ｄｃで示される３個の黒塗り丸印は、グループ４ａ，４ｂ，４ｃの重心を表す。同様にして、符号Ｄで示される黒塗り四角印は、テキスト集合４の重心を表す。

【００４８】図２から明らかなように、ファジイ集合化部（Fuzzy Clustering）５ａは、検索の結果得られたテキスト集合４に対して、公知のファジイクラスタリング処理を施すことにより、テキスト集合４を複数個（この例では３個）のグループ４ａ，４ｂ，４ｃに分割する。一方、主題分類情報生成部（Subject Generation）５ｂは、こうして得られた各グループ４ａ，４ｂ，４ｃのそれぞれについて、当該グループの属性を表現する主題分類情報Ｃａ，Ｃｂ，Ｃｃを生成する。図から明らかなように、このようにして得られる当該グループの属性を表現する主題分類情報Ｃａ，Ｃｂ，Ｃｃは、各グループ４ａ，４ｂ，４ｃの重心Ｄａ，Ｄｂ，Ｄｃとは異なるもの

であり、まさしくそれぞれのグループの属性を端的に表したものとなる。尚、これらのファジイ集合化部（Fuzzy Clustering）５ａ及び主題分類情報生成部（Subject Generation）５ｂの処理内容については、後に、更に詳しく説明する。

【００４９】図１に戻って、６は同様に本発明の要部に相当する検索結果提示部（Hit Windowと記す）であり、この検索結果提示部（Hit Window）６では、前述の経過により得られた情報（テキスト集合Ｂ、重心Ｄ、主題分類情報Ｃ）を所定の表示態様に加工したのち、ユーザ（Userと記す）７に対して提示する。それらの表示態様についても、後に詳細に説明する。

【００５０】尚、図１においては、実線により表された本発明による情報の流れと破線により表された従来装置による情報の流れとが同時に示されている。すなわち、従来装置にあっては、検索結果提示部（Hit Window）６では、破線矢印８に示されるように、検索の結果得られたテキスト集合Ｂをそのままユーザ７に提示するのみであり、この場合、テキスト集合Ｂに含まれるテキスト数が多量の場合、目的とするテキストを探し出すのにユーザは不便を来す。これに対して、本発明にあっては、検索結果提示部（Hit Window）６では、実線矢印９に示されるように、検索の結果得られたテキスト集合Ｂのみならず、各分類の重心（Clustercentroids）Ｄ並びに主題分類情報（Group Subject）Ｃまでもがユーザ７に提示されることとなるため、特に、この主題分類情報（Group Subject）Ｃを手掛かりとして、目的とするテキストを容易に探し出すことが可能となる。すなわち、実線矢印１０に示されるように、このようにして得られた主題分類情報Ｃ（図２のＣ１，Ｃ２，Ｃ３に相当する）をそのままフィードバック検索質問（FeedBack Query）Ｃとして検索質問入力部（Query Input）１に与えれば（図２の実線矢印１１に検索質問が分岐“Query Splitting”する様子を示す）、テキスト集合４を的確に絞込み、目的とするテキストを容易に探し出すことができ、すなわち高度な適応検索（relevance feedback）を行わせることができるのである。

【００５１】次に、以上概念的に説明したテキスト検索装置を、さらにその画面表示態様及びそれを実現するためのデータ処理を中心として、図３以下の図面を参照して詳細に説明する。

【００５２】本発明に係るテキスト検索装置におけるデータ処理の全体を図３のゼネラルフローチャートに示す。尚、このゼネラルフローチャートに示される処理は、所定のシステムメニューにおいて、そのメニュー項目のひとつを選択することにより起動される。

【００５３】同図において処理が開始されると、検索装置を構成する画像表示器の画面上には所定の表示態様により検索画面が表示される（ステップ３０１）。このようにして表示される検索画面の一例を図４に示す。同図に示されるように、表示画面は縦長長方形状のウィンドウＷ１により構成されており、その上部略３分の１の部分は検索質問入力領域Ａ１とされており、また下部略３分の２の部分は検索結果出力領域Ａ２とされている。

【００５４】検索質問入力領域Ａ１内には検索質問入力用のウィンドウＷ２が設けされており、このウィンドウＷ２の上側には、入力ガイド文（Enter Query in plain English）１２が、またその右側には、前述した検索エンジン（Search Engine）３に対する起動指令を与えるための起動ボタン（図中ＯＫと記す）１３と、検索質問（Query）を取り消すための取り消しボタン（図中ＣＡＮＣＥＬと記す）１４と、システムに対して操作支援等を求めるためのヘルプボタン（図中ＨＥＬＰと記す）１５とが設けられている。

【００５５】検索結果出力領域Ａ２内には検索結果出力用のウィンドウＷ３が設けられており、このウィンドウＷ３の右側にはスクロールバー１６が設けられている。更に、この検索結果出力領域Ａ２の下側には、検索結果としてテキスト全文出力を要求するための全文要求ボタン（図中Ｆｕｌｌ　Ｔｅｘｔと記す）１７と、ＱＢＥボタン１８と、検索結果の分類化を要求するための分類化要求ボタン（図中Ｇｒｏｕｐｉｎｇと記す）１９と、検索結果としてテキスト抄録出力を要求するための抄録要求ボタン（図中Ｓｕｍｍａｒｉｚｅと記す）２０と、画面を検索結果初期出力状態に戻すための復帰ボタン（図中Ｂａｃｋと記す）２１とが設けられている。

【００５６】尚、以上の各種のボタン１３，１４，１５，１６，１７，１８，１９，２０，２１の操作は、カーソルを希望のボタンに移動させた後、マウスのクリック操作等にて行われることは言うまでもない。

【００５７】そして、入力ガイド文（Enter Query in plain English）１２に従って、キーボードから検索質問を自然語（特に、この例では英語）にて、例えば、“I wantto know Clinton's political condition.”の如くに入力すると、この入力された検索質問２２はウィンドウＷ２内に表示されることとなる。

【００５８】この状態において、起動ボタン（図中ＯＫと記す）１３が操作されると、図３に戻って、検索／表示処理が実行され、検索エンジン（Search Engine）３が起動されて、検索質問に関連するテキスト集合４が文書データベース２より抽出され、この抽出されたテキスト集合の各構成テキストは検索質問２２との適合度の高い順にソートされ、そのタイトル２３のみがウィンドウＷ３内に表示される（ステップ３０２）。また、各テキストのタイトル２３の先頭部分には、当該テキストの検索質問に対する適合度を三段階（『高』、『中』、『低』）に区分して表す適合度マーク２４ａ，２４ｂ，２４ｃが表示される。ここで、黒色塗り潰しの丸印にて示される適合度マーク２４ａは適合度『高』に、灰色塗り潰しの丸印にて示される適合度マーク２４ｂは適合度

『中』に、白抜きの丸印にて示される適合度マーク２４ｃは適合度『低』にそれぞれ対応している。

【００５９】以後、図３に戻って、システム側においては文書処理機能の選択を待機する状態となる（ステップ３０３）。この状態において、図４の画面に示される分類化要求ボタン（Ｇｒｏｕｐｉｎｇ）１９が操作されると、本発明の要部である分類化処理が実行される（ステップ３０６）。

【００６０】分類化処理の詳細を図５に示す。同図において処理が開始されると、所定の案内画面を提示することにより、分類グループ数ｇの指定を待機する状態となる（ステップ５０１）。この状態において、分類グループ数ｇの指定（この例では『５』）が完了すると、本発明の特徴部分である文書特徴量の抽出処理（ステップ５０２）、ファジイ・クラスタリング処理（Fuzzy Clusteringと記す）（ステップ５０３）、及び主題分類情報の生成処理（ステップ５０４）が順に実行される。

【００６１】文書特徴量の抽出処理（ステップ５０２）では、次のようにして、文書抽象化と文書特徴ベクトルの生成が行われる。文書は重み付けられた語の集合（語を構成要素とするベクトル）によって表され、文書の集合は語を構成要素とする行列として表される。そのため、各文書の特徴となる単語（重要語）を自動的に切り出し、単語の種類を次元ｍとし、各要素が文書単位の単語の出現頻度に比例するようなベクトル表現Ｆｉを用いることによって、文書は数１の如くに抽象化される。

【００６２】

【数１】

$$Fi=(fi1, fi2, ..., fim)$$

Ｆｉ：文書ｉの特徴ベクトル
ｆｉｊ：単語ｊの文書ｉに対する重み
（頻度、或は他の評価値）

数１

文書ベクトル集合の例を表１に示す。この例では、文書集合の構成文書（Ｆ１，Ｆ２，Ｆ３…）のそれぞれに含まれる重要語（Clinton, Singapore, China…）の重み（例えば、頻度）が示されている。

【００６３】

【表１】

	Clinton	Singapore	China	...
F1	0.8	0.4	0.0	
F2	0.6	0.7	0.0	
F3	0.5	0.0	0.7	
...				

文書ベクトル集合の例

表１

表１に示される文書ベクトル集合を文書空間に展開した例を図６に示す。この例では、前述の重要語（Clinton, Singapore, China…）を座標軸とする文書空間に文書集合の各構成文書（Ｆ１，Ｆ２，Ｆ３…）が展開されている。

【００６４】続くファジイ・クラスタリング処理（ステップ５０３）では、検索結果としての文書集合の特徴行列に対し、公知のＦＣＭ法を用いてファジイ・クラスタリングを行うことにより、次の２種類の分類情報（Ｖｃ，Ｕｉ）が生成される。

【００６５】１）　各分類の代表文書特徴ベクトルＶｃ

【数２】

$$Vc=(vc1, vc2, ..., vcm)$$

Ｖｇ：分類グループｇの代表文書ベクトル
ｖｃｊ：単語ｊの分類グループｃの代表文書に対する重み

数２

２）　各文書の各分類カテゴリーへの所属度Ｕｉ

【数３】

$$Ui=(ui1, ui2, ..., uig)$$

Ｕｉ：文書の各分類グループへ所属度ベクトル
ｕｉｃ：文書ｉの分類グループｃへの所属度

数３

文書分類所属度の例を表２に示す。この例では、各文書の所属度（Ｕ１，Ｕ２，Ｕ３…）が各分類グループ（Ｇ１，Ｇ２，Ｇ３…）毎に示されている。

【００６６】

【表２】

	G1	G2 ... Gg
U1	0.7	0.2 ...
U2	0.8	0.1 ...
U3	0.3	0.6 ...
...		

文書分類所属度の例

表２

続く分類主題情報の生成処理（ステップ５０４）では、次の２種類の方式により、分類主題情報の生成が行われる。

【００６７】１）　キーワード方式

このキーワード方式は、各分類グループの主題を幾つかのキーワードの組み合わせにより表現する方式であり、その際に、キーワードの抽出には次の２種類の方式が考えられる。第１の方式は、該当分類の代表文書ベクトルＶｃにおける重みの高い要素の単語を順番にｋ個抽出してそれらの単語をそのグループの主題を表す情報として用いるものである。第２の方式は、該当分類の文書集合に対して所属度の高い順にｒ個の文書ベクトルを選出し、そのｒ個の文書ベクトル集合において出現文書数の高いものから順にｋ個の単語を抽出して、そのグループの主題情報を表す情報として用いるものである。

【００６８】２）　テキスト方式

このテキスト方式では、上記のキーワード方式で主題情報を生成するために選出されたｒ個の文書の先頭段落のテキスト（タイトルを含む）に対し、キーワード方式で得られたキーワード主題情報を利用して文単位で文字列照合によりそれらのキーワードを最も多く所有するテキストを抽出し、そのテキスト文をそのグループの主題情報として用いるものである。

【００６９】このようにして得られた各グループの主題情報、すなわち分類主題情報（前述のキーワード群又はタイトル文等）は、後述するように、所定の提示順番にてユーザに提示されることとなる。ここで、検索された文書ｉの検索質問に対する適合度をＲｉ、分類グループの検索式への適合度をＧＲｃとすると、両者間には数４の関係が成立する。

【００７０】

【数４】

$$GRc=\sum_{i}^{rc}Ri/rc$$

ｒｃ：グループｃに対して所属度の高い順に選出された文書数

Ｒｉ：グループｃに対して選出されたｒｃ個の文書集合内のｉ文書の適合度

数４

ここで、数４に示された、グループｃに対して所属度の高い順に選出された文書数ｒｃ（ｃ＝１，…，ｇ；ｇ：分類数）の求め方を図７のフローチャートに示す。同図において、処理が開始されると、ｒｃの初期化（ｒｃ＝０）を行ったのち（ステップ７０１）、文書ｉの所属度の行データＵｉに対して最大の所属度が求められ（ステップ７０２）、その最大値と対応しているグループｃのメンバ数ｒｃが加算され（ステップ７０３）、以上の処理（ステップ７０２，７０３）がｉを＋１づつ加算しつつ（ステップ７０４）、その加算値がｉ＝ｎ（文書数）となるまで（ステップ７０５ＹＥＳ）繰り返されて、その結果ｒｃの値が最終的に求められることとなる。

【００７１】このようにして、分類主題情報の生成（提示順番の決定を含む）が完了すると（ステップ５０４）、求められた主題分類情報を用いた検索結果の動的表示処理が開始される（ステップ５０５）。

【００７２】検索結果の動的表示処理の詳細を図８のフローチャートに示す。同図において処理が開始されると、検索装置を構成する画像表示器の画面上に設定された検索結果出力領域Ａ２は、図９又は図１０に示されるように、上下に２分割され、これにより主題分類情報表示用ウィンドウ（Subject Window）Ｗ４と検索結果出力用ウィンドウ（Hit Window）Ｗ５とが現れる。そして、主題分類情報表示用ウィンドウ（Subject Window）Ｗ４において、所定の表示態様により、各分類主題情報の提示が行われる（ステップ８０１）。前述したように、この各分類主題情報の提示は、キーワード方式とテキスト方式とで行われる。

【００７３】キーワード方式による表示画面の一例を図９に示す。尚、この例では、検索されたテキスト集合が５個の分類グループに分割されている。同図に示される

ように、主題分類情報表示用ウィンドウ（Subject Window）Ｗ４内には、その左縁部に沿うようにして、分類グループ番号『１』～分類グループ番号『５』に対応する５個のグループボタン２５～２９が上下一列に配置されており、それらのグループボタン２５～２９の右側には、当該分類グループの主題を的確に表すキーワード群３０～３４が配列されている。この例では、分類グループ番号『１』に対応するグループボタン２５の右側には、キーワード群３０として、“SINGAPORE;CANE;PUNISH;US”が表示されており、分類グループ番号『２』に対応するグループボタン２６の右側には、キーワード群３１として、“DALAILAMA;MEET;CHINA;TIBET”が表示されており、分類グループ番号『３』に対応するグループボタン２７の右側には、キーワード群３２として、“MEET;LEADER;GOVERNMENT;OFFICIAL”が表示されており、分類グループ番号『４』に対応するグループボタン２８の右側には、キーワード群３３として、“NIXON;NATION;SINGAPORE;DIRECTIVE”が表示されており、分類グループ番号『５』に対応するグループボタン２９の右側には、キーワード群３４として、“QUESTION;CHARACTER;PEOPLE;POLITICS”が表示されている。

【００７４】また、これらの主題分類情報は、先に求められた提示順番に従い、検索質問（Query）との適合度の高いものから順に配列されている。すなわち、この例では、分類グループ番号『１』にて象徴される主題が最も検索質問との適合度が高く、分類グループ番号『５』にて象徴される主題が最も検索質問との適合度が低いこととなる。従って、ユーザー７は主題分類情報表示用ウィンドウ（Subject Window）Ｗ４内の表示順番から、自分の探している情報に最も近い分類グループを容易に知ることができ、しかもそれぞれの内容を端的に表すキーワード群３０～３４の内容に基いて、各分類グループの主題を大まかに確認することができる。そして、後に詳しく説明するように、分類結果表示処理（ステップ８０２）を起動することにより、当初の検索質問に沿うようにして、検索絞り込みを効率よく行うことができる。

【００７５】テキスト方式による表示画面の一例を図１０に示す。尚、この例でも、検索されたテキスト集合が５個の分類グループに分割されている。同図に示されるように、主題分類情報表示用ウィンドウ（Subject Window）Ｗ４内には、その左縁部に沿うようにして、分類グループ番号『１』～分類グループ番号『５』に対応する５個のグループボタン２５～２９が上下一列に配置されており、それらのグループボタン２５～２９の右側には、当該分類グループの主題を的確に表す短いテキスト文３５～３９が配列されている。この例では、分類グループ番号『１』に対応するグループボタン２５の右側には、テキスト文３５として、“Clinton Protest Singapore Caning. Mulls Response”が表示されており、分類グループ番号『２』に対応するグループボタン２６の右側には、テキスト文３６として、“Clinton Meets With Dalai Lama”が表示されており、分類グループ番号『３』に対応するグループボタン２７の右側には、テキスト文３７として、“IndianLeader Meet Clinton”が表示されており、分類グループ番号『４』に対応するグループボタン２８の右側には、テキスト文３８として、“Nixon Had LivingWill”が表示されており、分類グループ番号『５』に対応するグループボタン２９の右側には、テキスト文３９として、“Clinton News Confers-Text”が表示されている。

【００７６】また、これらの主題分類情報についても、先に求められた提示順番に従い、検索質問（Query）との適合度の高いものから順に配列されている。すなわち、この例では、分類グループ番号『１』にて象徴される分類グループの主題が最も検索質問との適合度が高く、分類グループ番号『５』にて象徴される分類グループの主題が最も検索質問との適合度が低いこととなる。従って、ユーザー７は主題分類情報表示用ウィンドウ（Subject Window）Ｗ４内の表示順番から、自分の探している情報に最も近い分類グループを容易に知ることができ、しかもそれぞれの内容を端的に表すテキスト文３５～３９の内容に基いて、各分類グループの主題を大まかに確認することができる。そして、後に詳しく説明するように、分類結果表示処理（ステップ８０２）を起動することにより、当初の検索質問に沿うようにして、検索絞り込みを効率よく行うことができる。

【００７７】次に、先に説明したファジイ・クラスタリングにより得られた各文書の各分類グループへの所属度Ｕｉを用いた、検索結果の最終表示のための処理について詳細に説明する。尚、この例では、分類結果の最終表示のためには３種類の処理が用意されており、これらの処理は図９又は図１０に示される画面において、グループボタン２５～２９のいずれか一つを操作することにより起動される（ステップ８０２）。

【００７８】先に説明したように、本発明では検索結果としての文書集合の特徴行列に対し、ＦＣＭ法を用いてファジイ・クラスタリングを行い、それにより各文書の各分類カテゴリーへの所属度Ｕｉが求められている。今仮に、５個の文書（００１，００２，００３，００４，００５）が存在し、それらの文書のそれぞれについて３個の分類カテゴリ（カテゴリ１、カテゴリ２、カテゴリ３）のそれぞれに対する所属度が表３の通りであると想定する。

【００７９】

【表３】

21 22

文書番号	カテゴリ１	カテゴリ２	カテゴリ３
００１	０．５０	０．３０	０．２０
００２	０．６０	０．１０	０．３０
００３	０．１０	０．８０	０．１０
００４	０．２５	０．３４	０．４１
００５	０．３５	０．１０	０．５５

割り付けの説明のための数値例

表３

以上の前提の元に、ファジイ分類結果の３種類の表示処理（１）～（３）を説明する。

【００８０】（１）　各文書の所属度の上位ｋ個の分類カテゴリーへ割り付ける場合

この表示処理にあっては、各文書（００１～００５）は所属度の高いものから順に選ばれたｋ個の分類カテゴリに割り当てられる。例えば、ｋ＝１とすると（２値化方式）、文書（００１）については最大所属度０．５０であるカテゴリ１に、文書（００２）については最大所属度０．６０であるカテゴリ１に、文書（００３）については最大所属度０．８０であるカテゴリ２に、文書（００４）については最大所属度０．４１であるカテゴリ３に、文書（００５）については最大所属度０．５５であるカテゴリ３にそれぞれ割り付けられる。これを分類カテゴリ（Ｇ１，Ｇ２，Ｇ３）別に整理すると、

カテゴリＧ１＝（００１，００２）　　；Ｎ１＝２
カテゴリＧ２＝（００３）　　；Ｎ２＝１
カテゴリＧ３＝（００４，００５）　　；Ｎ３＝２

となり、分類グループＧ１に含まれる文書数Ｎ１は２個、分類グループＧ２に含まれる文書数Ｎ２は１個、分類グループ３に含まれる文書数Ｎ３は２個とされる。そして、このようにして各カテゴリに属することとされた文書が、後に詳細に説明するように、グループ番号の指定と共に検索結果出力用ウィンドウ（HitWindow）Ｗ５内に表示されることとなる。

【００８１】以上の表示処理（１）を実現するためのプログラムの一例を図１１に示す。同図において処理が開始されると、ｋ値の設定処理（ステップ１１０１）及びｉ，ｃ，Ｎｃの初期化処理（ステップ１１０２）を実行した後、文書ｉの所属度行データｉに対するソート処理（ステップ１１０３）、最大所属度データ値から順にｋ個のグループ番号を抽出する処理（ステップ１１０４）、及び該当するｋ個のグループに文書ｉを登録すると同時にメンバ数を加算する処理（ステップ１１０５）が、文書番号ｉがｎになるまで繰り返され（ステップ１１０６）、文書番号ｉがｎに達すると各グループ毎の文書割り付け結果を出力して処理が終了（ステップ１１０７）する。

【００８２】（２）　ある閾値α以上の所属度値を有する分類カテゴリに割り付ける場合

この表示処理にあっては、各文書（００１～００５）はある閾値α以上の所属度値を有する分類カテゴリに割り付けられる。ここで、αとしては、例えば１／ｇ（ｇ：分類数）とすることが考えられる。表３に示される例では、ｇ＝３、α＝０．３３となるため、文書（００１）については所属度値が０．３３以上であるカテゴリ１に、文書（００２）については同様な理由でカテゴリ１に、文書（００３）については同様な理由でカテゴリ２に、文書（００４）については同様な理由でカテゴリ２とカテゴリ３に、文書（００５）については同様な理由でカテゴリ１とカテゴリ３に割り付けられる。これを分類カテゴリ（Ｇ１，Ｇ２，Ｇ３）別に整理すると、

カテゴリＧ１＝（００１，００２，００５）　　；Ｎ１＝３
カテゴリＧ２＝（００３，００４）　　；Ｎ２＝２
カテゴリＧ３＝（００４，００５）　　；Ｎ３＝２

となり、分類グループＧ１に含まれる文書数Ｎ１は３個、分類グループＧ２に含まれる文書数Ｎ２は２個、分類グループ３に含まれる文書数Ｎ３は２個とされる。そして、このようにして各カテゴリに属することとされた文書が、後に詳細に説明するように、グループ番号の指定と共に検索結果出力用ウィンドウ（HitWindow）Ｗ５内に表示されることとなる。

【００８３】以上の表示処理（２）を実現するためのプログラムの一例を図１２に示す。同図において処理が開始されると、α値の設定処理（ステップ１２０１）及びｉ，ｃ，Ｎｃの初期化処理（ステップ１２０２）を実行した後、文書ｉの所属度行データｉに対するｕｉｃ＞α のグループ番号を抽出する処理（ステップ１２０３）、該当する各グループに文書ｉを登録すると同時にメンバ数を加算する処理（ステップ１２０４）が、文書番号ｉがｎになるまで繰り返され（ステップ１２０５）、文書番号ｉがｎに達すると各グループ毎の文書割り付け結果を出力して処理が終了（ステップ１２０６）する。

【００８４】（３）　カテゴリの確率分布を考慮して分類カテゴリに割り付ける場合

この表示処理にあっては、各文書（００１～００５）はカテゴリの確率分布を考慮して分類カテゴリに割り付けられる。ここで、文書の分類カテゴリーの確率分布（Ｐｃ）は数５に従って求められ、また分類ｃの文書数Ｎｃ

23

は数６に従って求められる。

【００８５】

【数５】

$$Pc = rc / n$$

rc：（１）の２値化方式により得られた分類ｃの文書数

n　：文書集合の全文書数

数５

【００８６】以上の表示処理（３）を実現するためのプログラムの一例を図１３に示す。同図において処理が開始されると、α値の設定処理（ステップ１３０１）、ｉ，ｃ，Ｎｃの初期化処理（ステップ１３０２）、文書の分類カテゴリーの確率分布（Ｐｃ＝ｒｃ／ｎ）を求める処理（ステップ１３０３）、分類ｃの文書数のＮｃを求める処理（ステップ１３０４）が順次に実行される。その後、文書ｃの所属度列データｕｉｃに対するソート処理（ステップ１３０５）、最大所属度値から順に対応のＮｃ個のメンバの文書番号を抽出する処理（ステップ１３０６）、及び該当のグループｃにＮｃ個の文書を登録する処理（ステップ１３０７）が、分類ｃが分類数ｇになるまで繰り返され（ステップ１３０８ＮＯ）、分類ｃが分類数ｇに達すると（ステップ１３０８ＹＥＳ）、各グループ毎の文書割り付け結果を出力して処理が終了する（ステップ１３０９）。

【００８７】次に、以上説明した３種類の割り付け処理（１）～（３）のいずれかにて各分類グループに割り付けられた文書が、表示画面上の検索結果出力用ウィンドウ（Hit Window）Ｗ５内にどのような態様で表示されるかを説明する。

【００８８】図９に示される画面上において、いずれかのグループボタン（この例では、グループボタン２６）が指定操作されると、上述した３種類の割り付け処理

24

【数６】

数６

表３に示される例では、Ｐ１＝０．４、Ｐ２＝０．２、Ｐ３＝０．４となり、またＮ（０．３３）＝７となるため、Ｎ１＝２．８（約３）、Ｎ２＝１．４（約１）、Ｎ＊３＝２．８（約３）となる。これを分類カテゴリ（Ｇ１，Ｇ２，Ｇ３）別に整理すると、

カテゴリＧ１＝（００１，００２，００５）　；Ｎ１＝２

カテゴリＧ２＝（００３）　；Ｎ２＝１

カテゴリＧ３＝（００２，００４，００５）　；Ｎ３＝２

となる。そして、このようにして各カテゴリに属することとされた文書が、後に詳細に説明するように、グループ番号の指定と共に検索結果出力用ウィンドウ（Hit Window）Ｗ５内に表示されることとなる。

（１）～（３）のいずれかにて各分類グループに割り付けられた文書に相当する短いテキスト文（この例ではタイトル等を含む当該テキストの先頭部分）４０～４４が、検索結果出力用ウィンドウ（Hit Window）Ｗ５内に表示されることとなる（ステップ８０２）。

【００８９】すなわち、この例では、キーワード群３１（"DALAILAMA;MEET;CHINA;TIBET"）にて象徴化される分類グループ番号『２』が指定されたことにより、検索結果出力用ウィンドウ（Hit Window）Ｗ５内には、これに関連する５個のテキスト文４０（"Clinton Meets With Dalai Lama）、テキスト文４１（"Clinton, Gore MeetDalai Lama on Tibetan Right）、テキスト文４２（"China Warns Clinton Notto Meet Dalai Lama"）、テキスト文４３（"Clinton May Meet Dalai Lama before China Decision"）、テキスト文４４（"Indian Leader Meet Clinton"）が表示されている。しかも、これらのテキスト文４０～４４は、図中『Ｇ』と記されたグループ適合度順指定ボタン５１が操作されていることから、当該指定された分類グループ番号『２』で象徴化される分類グループとの適合度の順に配列して表示されている。尚、符号４５，４６はそれぞれその左側に位置するウィンドウＷ４，Ｗ５のスクロールバー、４９は分類グループ数の表示である。

【００９０】更に、検索結果出力用ウィンドウ（Hit Window）Ｗ５内において、各テキスト文４０～４４のそれぞれの先頭部分には、各テキスト文４０～４４が当該分類グループに対して有する適合度を３段階に表す３種類の適合度マーク（４７ａ，４７ｂ，４７ｃ）と、各テキ

25

スト文４０～４４が当該検索質問２２に対して有する適合度を３段階に表す３種類の適合度マーク（４８ａ，４８ｂ，４８ｃ）が表示されている。この例では、当該分類グループとの適合度を表す適合度マーク（４７ａ，４７ｂ，４７ｃ）は基本形状が雪印であり、適合度『高』に相当する適合度マーク４７ａについてはその中心の小円形部分を黒色塗り潰しに、適合度『中』に相当する適合度マーク４７ｂについてはその中心の小円形部分を灰色塗り潰しに、更に適合度『低』に相当する適合度マーク４７ｃについてはその中心の小円形部分を白抜きとしている。また、当該検索質問との適合度を表す適合度マーク（４８ａ，４８ｂ，４８ｃ）は基本形状が丸印であり、適合度『高』に相当する適合度マーク４８ａについては黒色塗り潰しに、適合度『中』に相当する適合度マーク４８ｂについては灰色塗り潰しに、更に適合度『低』に相当する適合度マーク４８ｃについては白抜きとしている。

【００９１】従って、この検索結果出力用ウィンドウ（Hit Window）Ｗ５内の表示内容４０～４４により、ユーザー７は検索結果であるテキスト集合の中で分類グループ番号『２』のグループに属するテキスト集合を、適合度マーク（４７ａ，４７ｂ，４７ｃ）を頼りとして、該分類グループ『２』との適合度の高いものから順に確認しつつ、目的とする情報を的確に見つけ出すことができる。加えて、適合度マーク（４８ａ，４８ｂ，４８ｃ）を参照することにより、各テキスト文４０～４４と検索質問２２との適合度も知ることができるため、双方のマーク４７，４８を参考として、一層確実な検索絞り込みを行うことができる。尚、図示されてはいないが、図中『Ｒ』と記された検索質問適合度順指定ボタン５０が操作された場合には、図８において分類主題表示指定処理（ステップ８０４）が実行されて、各テキスト文４０～４４は当該検索質問２２との適合度の順に配列されて表示されることとなる。従って、検索質問適合度順指定ボタン５０とグループ適合度順指定ボタン５１とのいずれを選択するかにより、各テキスト文４０～４４の配列を変更しつつ、検索結果を所望の検索方向に沿って確認することができる。

【００９２】一方、例えば図９に示される検索結果が表示されている状態において、操作支援要求ボタン（ＨＥＬＰ）１５が操作されると、図８に戻って、主題表示オプション処理（ステップ８０５）が実行され、主題分類情報表示用ウィンドウ（Subject Window）Ｗ４内の表示は、図１０に示されるように、前述のキーワード方式からテキスト方式へと切り替わる。そのため、キーワード方式では当該分類グループの内容が把握しにくい場合でも、このテキスト方式による主題分類情報の表示によれば、当該分類グループにて象徴化される主題をより的確に知ることができる。尚、各ウィンドウＷ４，Ｗ５内に表示データが収まらない場合には、スクロールバー４５，４６の操作にて表示内容をスクロールしつつ確認できることは言うまでもない。

26

【００９３】

【発明の効果】以上の説明で明らかなように、この発明によれば、文書検索結果に対する確認を容易として、検索効率の向上、並びに、検索漏れの防止による検索精度の向上を図ることができ、しかも、提示された主題情報がデータを如何に効率的に絞り込めるかの指針にもなり、この付加された応答情報を利用して高度な適応検索（Relevance Feedback）を行わせることができる。

【図面の簡単な説明】

【図１】本発明方法及び装置が適用されたテキスト検索装置の構成を概念的に示すブロック図である。

【図２】ファジイ集合化部（Fuzzy Clustering）及び主題分類情報生成部（Subject Generation）の作用を概念的に示す説明図である。

【図３】本発明に係るテキスト検索装置の動作の全体を概略的に示すゼネラルフローチャートである。

【図４】本発明に係るテキスト検索装置においてグループ化処理を伴わない検索動作を実行させた状態を示す画面説明図である。

【図５】本発明に係るテキスト検索装置における主題分類情報の生成処理を中心として示すフローチャートである。

【図６】本発明に係るテキスト検索装置における文書抽象化と文書ベクトルの生成を概念的に示す説明図である。

【図７】本発明に係るテキスト検索装置におけるグループｃのメンバ数ｒｃを求めるための処理を示すフローチャートである。

【図８】本発明に係るテキスト検索装置における主題分類情報による検索結果の動的処理を示すフローチャートである。

【図９】本発明に係るテキスト検索装置においてグループ化処理を伴なう検索動作をキーワード方式にて実行させた状態を示す画面説明図である。

【図１０】本発明に係るテキスト検索装置においてグループ化処理を伴なう検索動作をテキスト方式にて実行させた状態を示す画面説明図である。

【図１１】本発明に係るテキスト検索装置にて検索結果をグループ別に表示するにおいて、各文書の所属度の上位ｋ個の分類カテゴリーへの割り付け処理を示すフローチャートである。

【図１２】本発明に係るテキスト検索装置にて検索結果をグループ別に表示するにおいて、α値以上の所属度値をもつ分類カテゴリーへの割り付け処理を示すフローチャートである。

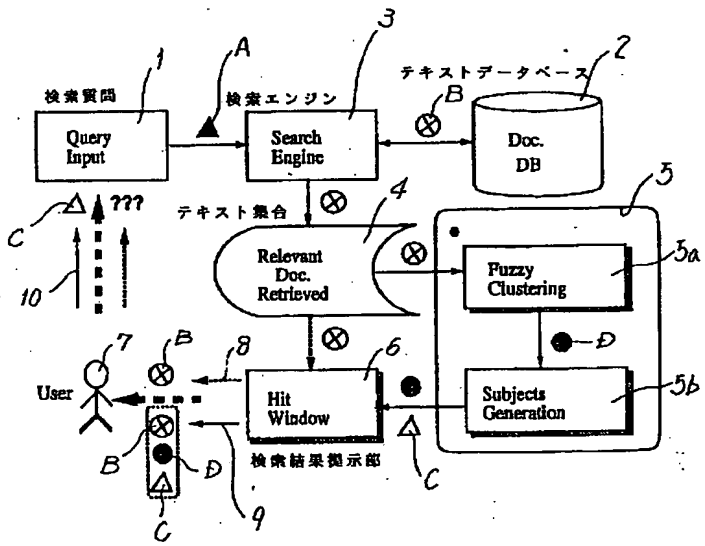
【図１３】本発明に係るテキスト検索装置にて検索結果をグループ別に表示するにおいて、カテゴリーの確率分布を考慮した分類カテゴリーへの割り値付け処理を示す

フローチャートである。

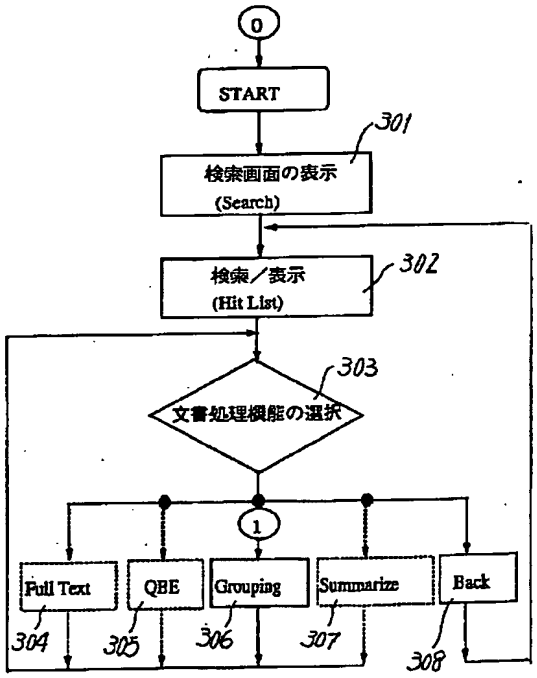
【符号の説明】

1　検索質問入力部
2　文書データベース
3　検索エンジン
4　抽出された関連テキスト集合
4a，4b，4c　分類グループ
5　加工処理部
5a　ファジイ集合化部
5b　主題分類情報生成部
6　検索結果提示部
7　ユーザー
12　入力ガイド文
13　起動ボタン
14　取り消しボタン
15　ヘルプボタン
16　スクロールバー
17　全文要求ボタン
18　QBEボタン
19　分類化要求ボタン
20　抄録要求ボタン
21　復帰ボタン
22　検索質問
23　テキスト集合を構成する各テキストのタイトル
24a，24b，24c　適合度マーク
25～29　グループボタン
30～34　キーワード群
35～39　テキスト文
40～44　テキスト文
45，46　スクロールバー
49　分類グループ数の表示
47a，47b，47c　グループ毎の適合度マーク
48a，48b，48c　検索質問に対する適合度マーク
49　分類グループ数の表示
50　検索質問適合度順指定ボタン
51　グループ適合度順指定ボタン
A　オリジナル検索質問
B　抽出された各構成テキスト
Ca，Cb，Cc　主題分類情報
Da，Db，Dc　グループの重心
A1　検索質問入力領域
A2　検索結果出力領域
W2　検索質問入力用のウィンドウ
W3　検索結果出力用のウィンドウ
W4　主題分類情報表示用ウィンドウ
W5　検索結果出力用ウィンドウ

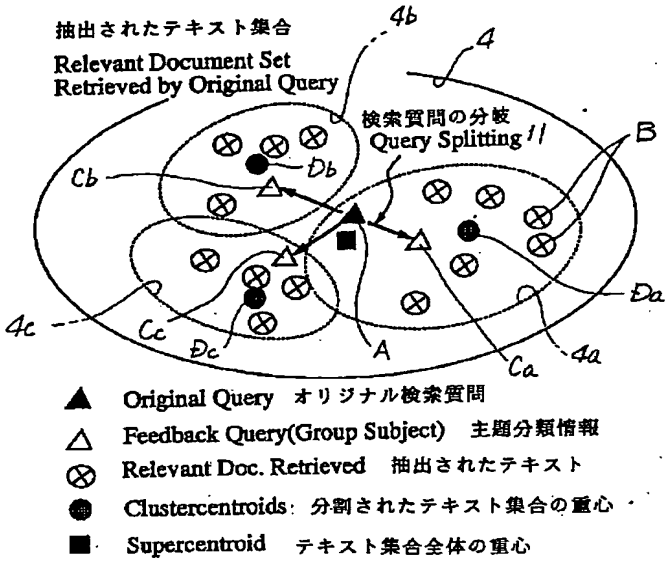
【図1】



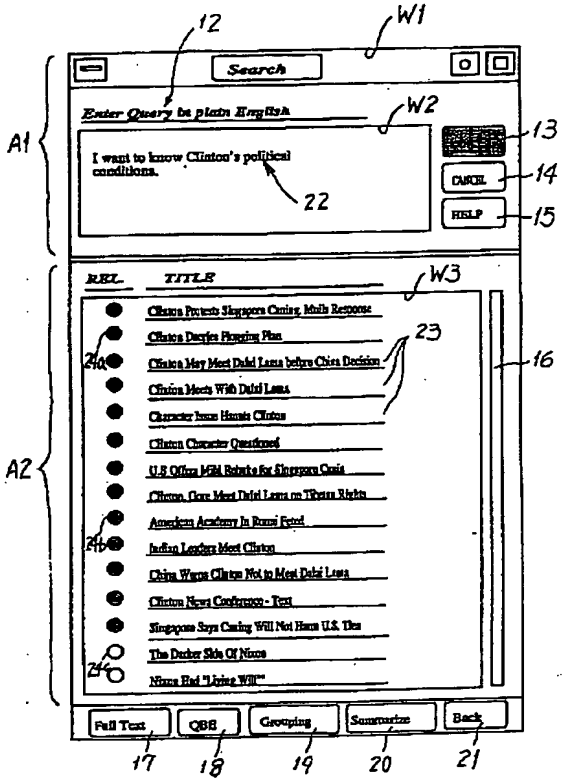
【図3】



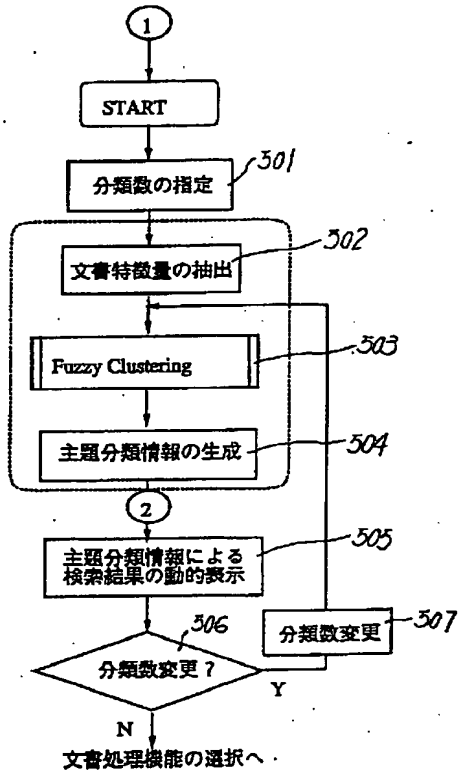
【図２】



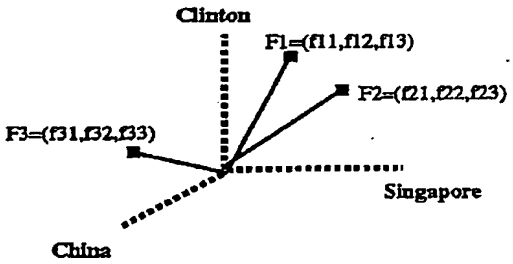
【図４】



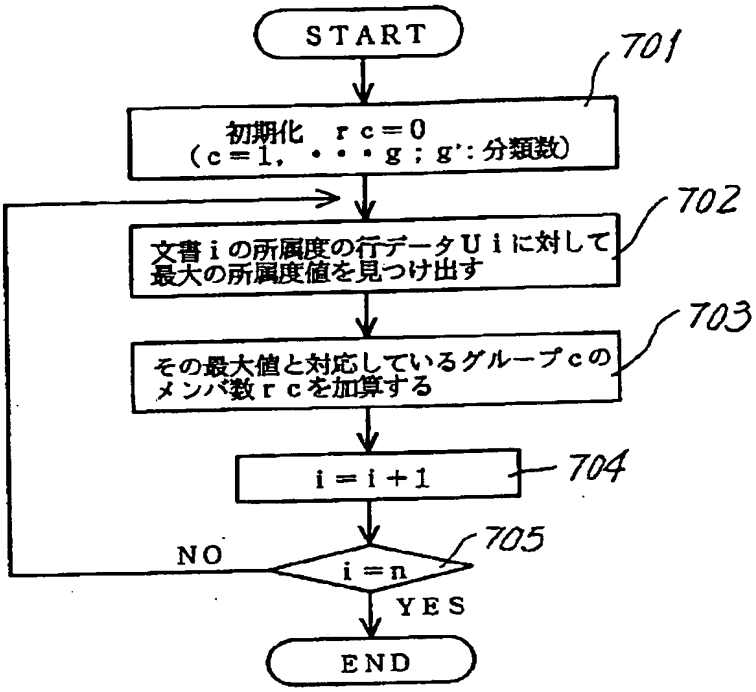
【図５】



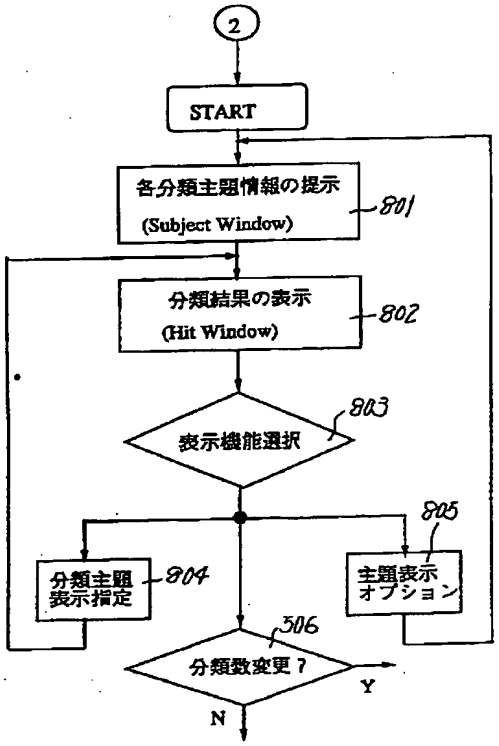
【図６】



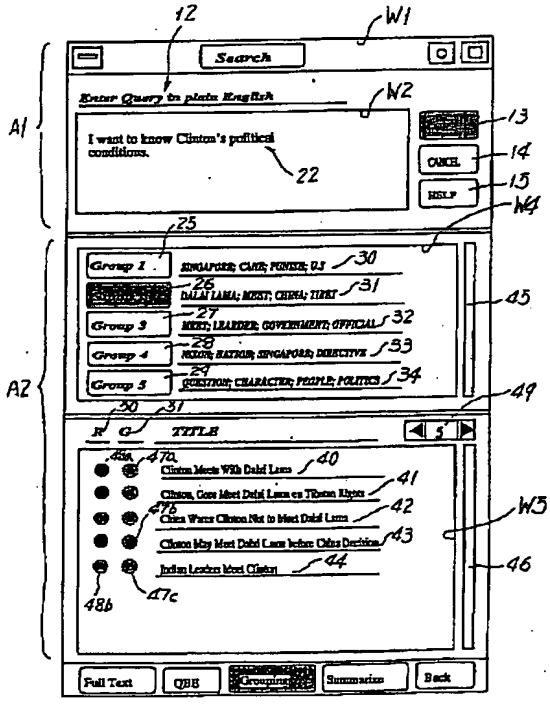
【図７】



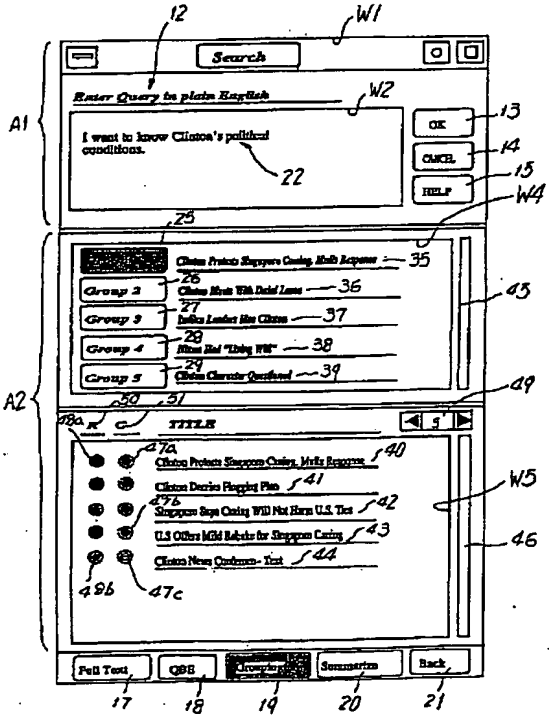
【図８】



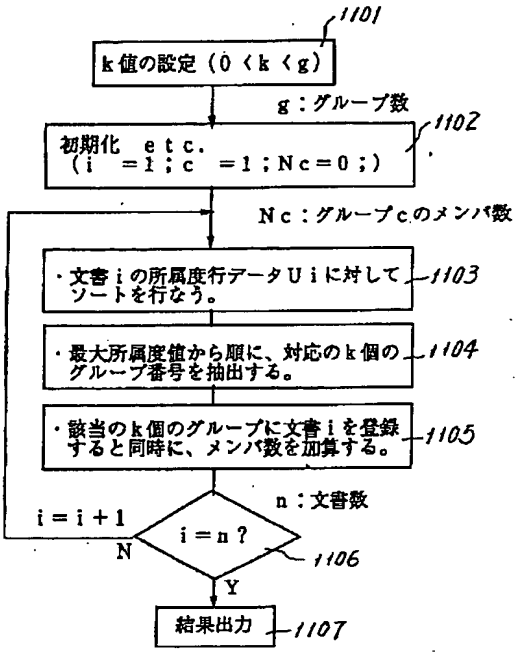
【図９】



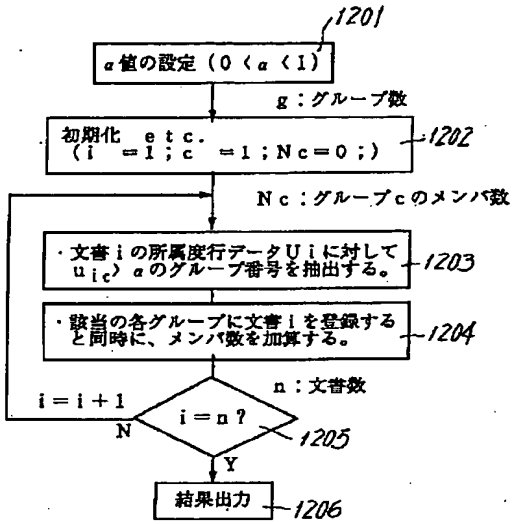
【図１０】



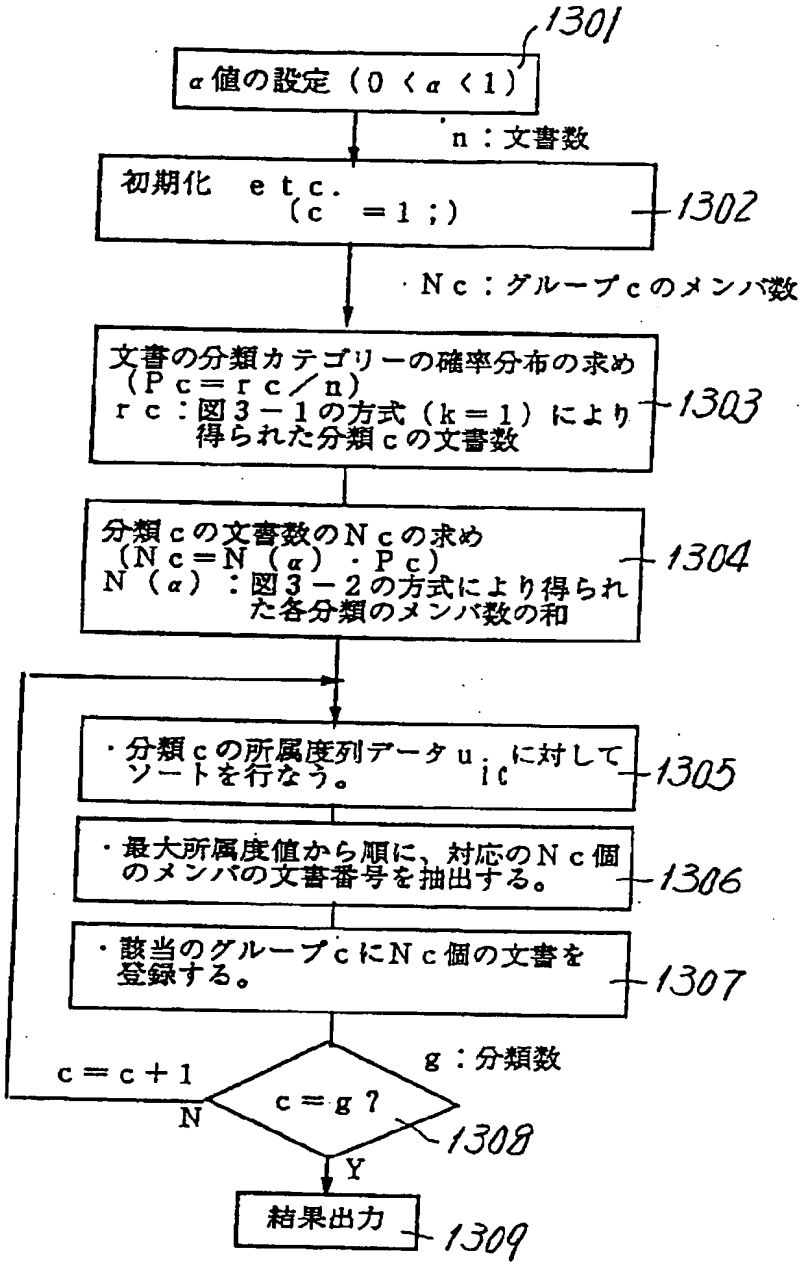
【図１１】


1101
k値の設定（0＜k＜g）
g：グループ数
1102
初期化　etc.
（i　＝1；c　＝1；Nc＝0；）
Nc：グループcのメンバ数
・文書iの所属度行データUiに対してソートを行なう。
1103
・最大所属度値から順に、対応のk個のグループ番号を抽出する。
1104
・該当のk個のグループに文書iを登録すると同時に、メンバ数を加算する。
1105
n：文書数
i＝i＋1
i＝n？
N
Y
1106
結果出力
1107

【図１２】


1201
α値の設定（0＜α＜1）
g：グループ数
初期化　etc.
（i　＝1；c　＝1；Nc＝0；）
1202
Nc：グループcのメンバ数
・文書iの所属度行データUiに対して
uic＞αのグループ番号を抽出する。
1203
・該当の各グループに文書iを登録すると同時に、メンバ数を加算する。
1204
n：文書数
i＝i＋1
i＝n？
N
Y
1205
結果出力
1206

【図１３】


1301
α値の設定（0＜α＜1）
n：文書数
初期化　etc.
（c　＝1；）
1302
・Nc：グループcのメンバ数
文書の分類カテゴリーの確率分布の求め
（Pc＝rc／n）
rc：図3－1の方式（k＝1）により得られた分類cの文書数
1303
分類cの文書数のNcの求め
（Nc＝N（α）・Pc）
N（α）：図3－2の方式により得られた各分類のメンバ数の和
1304
・分類cの所属度列データu_{ic}に対してソートを行なう。
1305
・最大所属度値から順に、対応のNc個のメンバの文書番号を抽出する。
1306
・該当のグループcにNc個の文書を登録する。
1307
g：分類数
c＝c＋1
c＝g？
N
Y
1308
結果出力
1309